

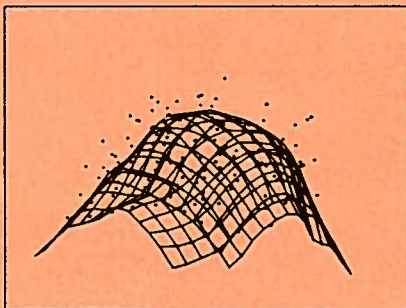
# Local Likelihood Estimation

*Robert Tibshirani*

**Technical Report No. 4**

**September 1984**

**Laboratory for  
Computational  
Statistics**



**Department of Statistics  
Stanford University**

# LOCAL LIKELIHOOD ESTIMATION

Robert J. Tibshirani

Department of Statistics

Stanford University

and

Computation Research Group

Stanford Linear Accelerator Center

## Abstract

Given scatterplot data  $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$ ,  $Y$  being a response and  $X$  a predictor, a *scatterplot smoother* uses *local averaging* to estimate the dependence of  $Y$  on  $X$ . A simple example is the *running lines smoother*, which fits a least squares line to the  $Y$  values falling in a window around each  $X$  value. A smoother generalizes the least squares line, which assumes the dependence of  $Y$  on  $X$  is linear.

In this paper, we extend the idea of local averaging to likelihood based models. One such application is to the class of *generalized linear models* (Nelder and Wedderburn (1972)). We enlarge this class by replacing the covariate form  $x\beta$  with an unspecified smooth function  $s(x)$ . This function is estimated from the data by a technique we call "*Local Likelihood Estimation*"—a type of local averaging. Multiple covariates are incorporated through a forward stepwise algorithm.

The main application that we discuss, however, is to the proportional hazards model of Cox (1972), for censored data. The proportional hazards assumption  $\lambda(t | x) = \lambda_0(t) \exp(x\beta)$  is replaced by  $\lambda(t | x) = \lambda_0(t) \exp(s(x))$ , and the function  $s(x)$  is estimated from the data by local likelihood estimation.

In a number of real data examples, the local likelihood technique proves to be effective in uncovering non-linear dependencies.

Finally, we give some asymptotic results for local likelihood estimates and provide some methods for inference.

**Keywords:** *smoothing, proportional hazards model, generalized linear models*

Work supported by the Department of Energy under contracts DE-AC03-76SF00515 and DE-AT03-81-ER10843, and by the Office of Naval Research under contract ONR N00014-81-K-0340, and by the U.S. Army Research Office under contract DAAG29-82-K-0056.

# LOCAL LIKELIHOOD ESTIMATION

Robert J. Tibshirani

## 1. Introduction.

Figure (1) contains 100 data pairs along with the least squares line summarizing the relationship of a response (say  $Y$ ) and a covariate ( $X$ ). In Figure (2), the least squares line has been replaced by a “scatterplot smooth.” This smooth was computed by a type of local averaging—around each  $X$  value a window of 20 points was formed and a least squares line was fit to the points in the window. The value of the smooth at  $X$  is given by the value of the “local line” at  $X$ . As we can see, the smooth captures the trend of the data better than the least squares line. The reason is simple—the smooth doesn’t make a rigid assumption about the form of the relationship between  $Y$  and  $X$ .

In recent years, there has been a great deal of interest in scatterplot smoothing by local averaging (see for example Cleveland(1979) and Friedman and Stuetzle(1981)) and the availability of fast computers has been essential in this development. These smooths are useful as a descriptive tool (as we have seen above) and also as building blocks for non-parametric regression models. Important developments in the latter area can be found in Friedman and Stuetzle (1981) and Breiman and Friedman(1982).

In this paper, we explore an application of smoothing ideas to other kinds of data. In particular, we consider  $(X, Y)$  data whose relationship is expressible through a likelihood function. Take for example the situation in which  $Y$  is a 0-1 response and  $X$  is a covariate. For such a data set, Figure (3) shows the logistic regression line, estimated by maximum likelihood. On the same plot, the observed logits are shown. (Since we can’t take the logit of 0 or 1, the  $Y$ ’s were grouped first). In Figure (4), the line has been replaced by a smooth. As was the case in the scatterplot example, the smooth does a better job of capturing the relationship between  $Y$  and  $X$  than the line does. In Figures (5) and (6), we see another example. Figure (5) shows the estimated log relative risk line given by Cox’s proportional hazards model, applied to a set of survival data. In Figure (6), the line has been replaced by a “log relative risk smooth”.

The smooths in Figures (4) and (6) were obtained from a procedure we call “local likelihood” estimation. The basic idea is simple extension of the local averaging technique used in scatterplot smoothing. Given a global method for estimating a linear response (e.g. maximum likelihood estimation in the linear logistic model), we apply it locally, estimating a separate line in a window around each  $X$  value. The value of the estimated line at  $X$  is the estimate of the smooth response function at  $X$ .

By varying the window size, we can control the smoothness of the estimated function. The larger the windows, the smoother the estimated function. When each window contains 100% of the data, the local likelihood procedure corresponds exactly to the global linear method.

Hence local likelihood completely generalizes linear likelihood estimation.

This paper is devoted to the study of local likelihood. We describe the method in general, showing how smooths like those in Figures (4) and (6) are obtained, and we will study some of its theoretical properties. In the exponential family, the local likelihood method extends the class of *generalized linear models* (Nelder and Wedderburn (1972)) by allowing covariates to enter the link function in a non-linear fashion. We investigate the linear logistic model, a member of this class, and its extension. We also explore in depth the application of the method to the proportional hazards model. This model was the motivating example behind local likelihood.

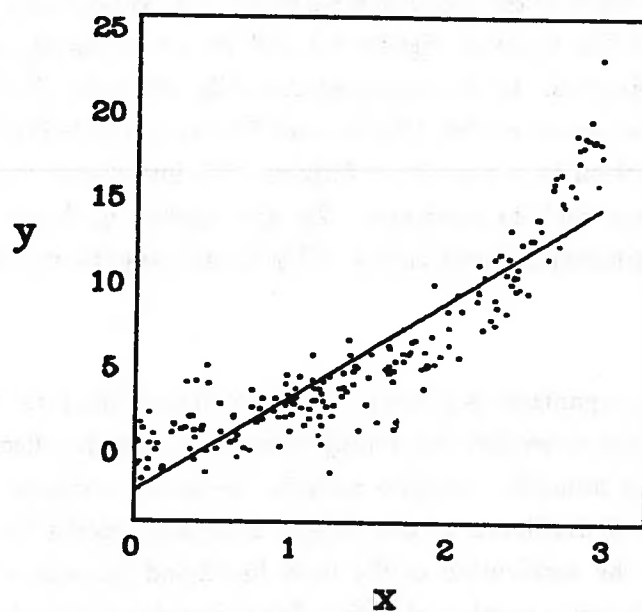
The sections are organized as follows. Section 2 defines the local likelihood method and describes the estimation procedure for a single covariate. We also discuss a forward stepwise algorithm for building multiple covariate models. Section 3 contains a short description of the application of local likelihood to the logistic regression model for binary data. Section 4 describes in detail the application of the local likelihood procedure to Cox's proportional hazards model. We discuss a number of topics: bootstrapping the models, robustifying the fit, and bias of the procedure. Finally, some real data examples are given.

Section 5 provides some asymptotic results for local likelihood estimates in the exponential family. Consistency and efficiency of the estimates are discussed. We conjecture (without proof) similar results for the proportional hazards model.

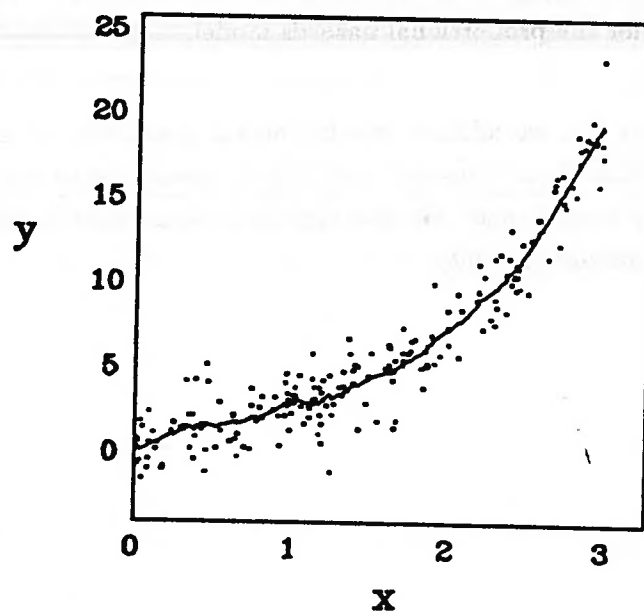
In the last section (6), we address two important questions: 1) how many parameters are used up by a local likelihood smooth? and 2) is it reasonable to use Akaike's Information Criterion to choose the window size? We give approximate answers to these questions, backing up our claims with a simulation study.

**4 Section 1: Introduction**

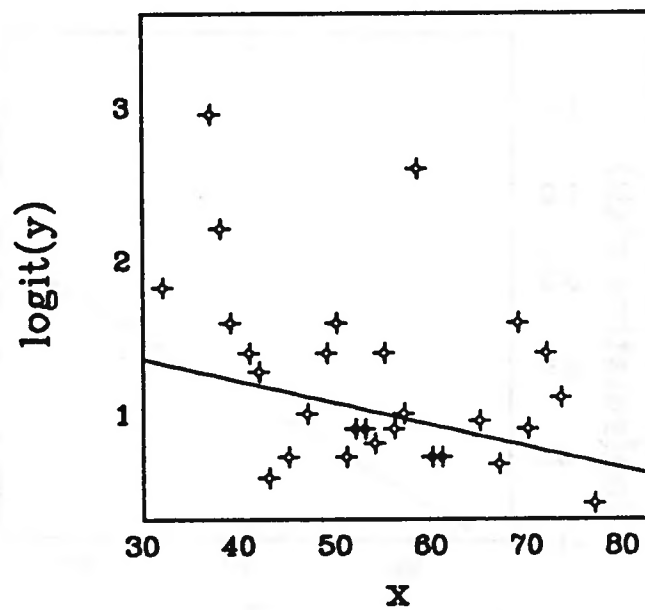
**Figure (1)**  
**Least Squares Line**



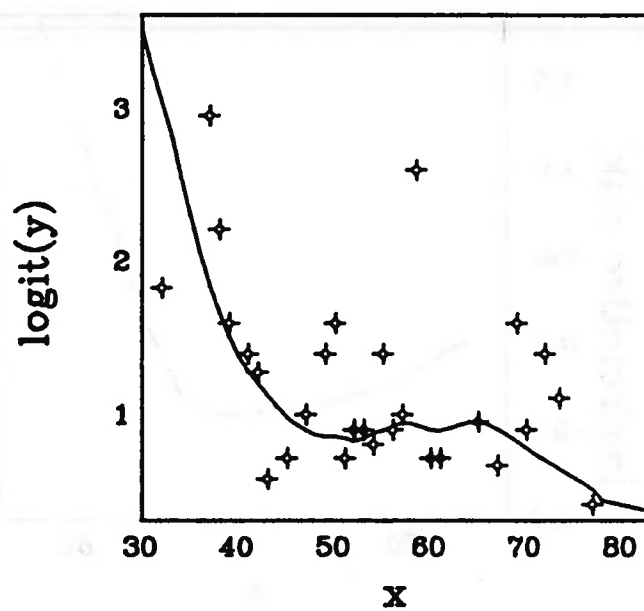
**Figure (2)**  
**Scatterplot Smooth**



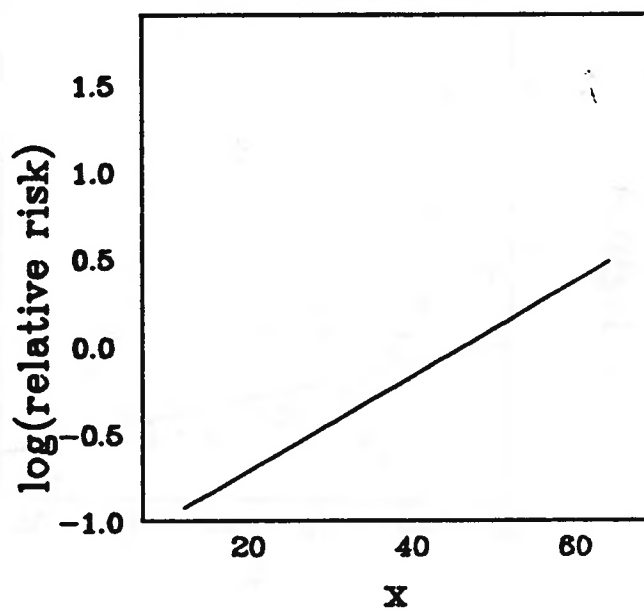
**Figure (3)**  
Logistic Line



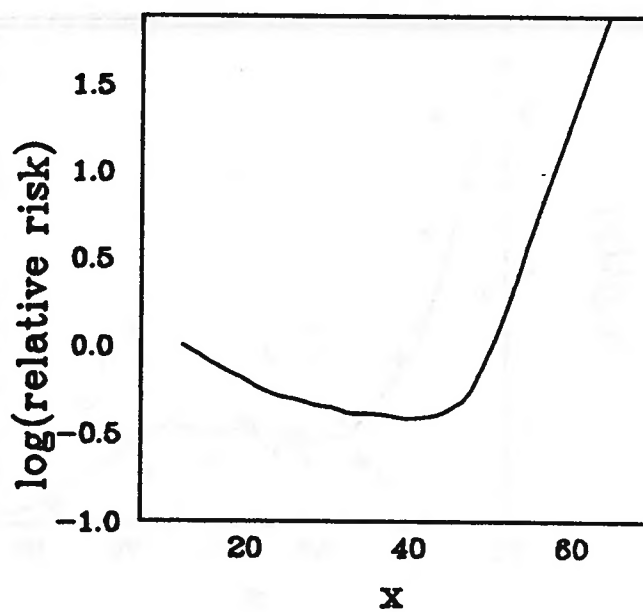
**Figure (4)**  
Local Likelihood Logistic Smooth



**Figure (5)**  
Relative Risk Line



**Figure (6)**  
Local Likelihood Relative Risk Smooth



## 2. Local Likelihood— A description.

### 2.1. Introduction

In this section we introduce the local likelihood idea. Since local likelihood estimation is a generalization of scatterplot smoothing, we begin with a review of the latter.

### 2.2. A Review of Scatterplot Smoothing

Given independent data pairs  $\{(x_1, y_1), \dots, (x_n, y_n)\}$ , assumed to be realizations of a response variable  $Y$  and a predictor  $X$ , a *scatterplot smoother* produces a decomposition of the form

$$y_i = s(x_i) + \epsilon_i \quad (1)$$

Here  $s(\cdot)$  is a “smooth” function and  $\epsilon_i$  is a residual error. We won’t define exactly what “smooth” means here; vaguely speaking, we’re thinking of  $s(\cdot)$  as a function less smooth than a straight line but smoother than an interpolating polynomial.

There are many ways to estimate  $s(\cdot)$ — we’ll concentrate here on the method of “local averaging”. It is motivated as follows. If we knew the joint distribution of  $Y$  and  $X$ , a reasonable way to find  $s(\cdot)$  would be to minimize  $E(Y - s(X))^2$ , where the expectation is taken over this joint distribution. Conditioning on  $X = x$ , this has solution  $\hat{s}(x) = E(Y | X = x)$  for each  $x$ . In practice, we don’t know this joint distribution but have only a sample from it. The idea, then, is to estimate  $E(Y | X = x)$  from the data. This leads to the class of *local average* estimates for  $s(\cdot)$ :

$$\hat{s}(x_i) = \mathbf{Ave}_{j \in N_i} s(x_j) \quad (2)$$

where “**Ave**” represents some averaging operator like the mean and  $N_i$  is a “neighborhood” of  $x_i$  (a set of indices of points whose  $x$  values are “close” to  $x_i$ ). The only type of neighborhoods we’ll consider in this paper are *symmetric nearest neighborhoods*. Assuming that the data points are sorted by increasing  $x$  value, these are defined by:

$$N_i = \{\max(i - \frac{k-1}{2}, 1), \dots, i-1, i, i+1, \dots, \min(i + \frac{k-1}{2}, n)\} \quad (3)$$

The parameter  $k$  is called the *span* of the smoother and controls the smoothness of the resulting estimate. The value of  $k$  must be chosen in some way from the data.

If **Ave** stands for arithmetic mean, then  $\hat{s}(\cdot)$  is the *running mean*, the simplest possible scatterplot smoother. The running mean is not a satisfactory smoother because it creates large biases at the endpoints and doesn’t reproduce straight lines (i.e. if the data lie exactly along a straight line, the smooth of the data will not be a straight line). A slight refinement of

## 8 Section 2: Local Likelihood—A description

the running average, the *running lines smoother* alleviates these problems. The running lines estimate is defined by

$$\hat{s}(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i}x_i \quad (4)$$

where  $\hat{\beta}_{0i}$  and  $\hat{\beta}_{1i}$  are the least squares estimates for the data points in  $N_i$ :

$$\begin{aligned} \hat{\beta}_{1i} &= \frac{\sum_{j \in N_i} (x_j - \bar{x}_i)y_j}{\sum_{j \in N_i} (x_j - \bar{x}_i)^2} \\ \hat{\beta}_{0i} &= \bar{y}_i - \hat{\beta}_{1i}\bar{x}_i \end{aligned} \quad (5)$$

and  $\bar{x}_i = \frac{1}{n} \sum_{j \in N_i} x_j$ ,  $\bar{y}_i = \frac{1}{n} \sum_{j \in N_i} y_j$ .

The running lines smooth is the most obvious generalization of the least squares line. When every neighborhood contains 100% of the data points, the smooth agrees exactly with the least squares line. For smaller spans, it produces less smooth estimates. Although very simple in nature, the running lines smoother produces reasonable results and has the advantage that the estimates can be updated. That is, to find  $\hat{s}(x_{i+1})$  from  $\hat{s}(x_i)$ , only a  $O(1)$  operation is needed. This reduces the overall algorithm from  $O(n^2)$  to  $O(n)$ .

### 2.3. Local Gaussian Smoothing

Since least squares estimation corresponds to maximum likelihood when the data are Gaussian, it is not surprising that the running lines smoother can be described as a “running maximum likelihood” method for Gaussian data. Assume as before that

$$y_i = s(x_i) + \epsilon_i \quad (6)$$

and in addition that  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ . Then for  $x$  in a neighborhood  $N_i$  of  $x_i$ , a reasonable approximation to  $s(x)$  is

$$s(x) \approx \beta_{0i} + \beta_{1i}x \quad (7)$$

Considering only the points in  $N_i$ , the maximum likelihood estimates of  $\beta_{0i}$  and  $\beta_{1i}$  are given by (5). Based on (7), this gives as an estimate of  $s(x_i)$ :

$$\hat{s}(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i}x_i \quad (8)$$

Hence running lines smoothing corresponds to finding approximate maximum likelihood estimates in a neighborhood around each data point.

We call this type of estimation “*LOCAL LIKELIHOOD ESTIMATION*” or “*LOCAL LIKELIHOOD*” for short. In this paper, we extend the idea of local likelihood to non-Gaussian likelihoods. It can be applied in principal to any situation in which the effect of a covariate is modelled through a likelihood. In fact, as we will see in the proportional hazards model, the “likelihood” doesn’t even have to be a likelihood in the strict sense.

## 2.4. Local Likelihood: General Definition

Suppose we have  $n$  data tuples of the form  $(y_i, x_i, \mathbf{c}_i)$ , where  $y$  is a response variable,  $x$  is a covariate or predictor variable, and  $\mathbf{c}$  is a vector containing any additional information. (In censored data problems,  $\mathbf{c}$  would indicate whether  $y$  is censored; in many problems (like regression),  $\mathbf{c}$  is empty.) Suppose that modelling considerations lead to maximization of a function of the form

$$L(\beta_0, \beta_1) = g(y_1, y_2, \dots, y_n, \theta_1, \theta_2, \dots, \theta_n, \mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n) \quad (9)$$

where  $\theta_i = \beta_0 + \beta_1 x_i$ . For example,  $L(\beta_0, \beta_1)$  could be a likelihood function and the estimates maximizing  $L(\beta_0, \beta_1)$  would be the maximum likelihood estimates. The *LOCAL LIKELIHOOD* method replaces  $\beta_0 + \beta_1 x_i$  with an arbitrary smooth function  $s(x_i)$ :

$$L(s(x_1), s(x_2), \dots, s(x_n)) = g(y_1, y_2, \dots, y_n, \theta_1, \theta_2, \dots, \theta_n, \mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n) \quad (10)$$

where  $\theta_i = s(x_i)$ . The problem is to estimate  $s(\cdot)$  at the points  $\{x_1, x_2, \dots, x_n\}$ . Maximization of  $L(s(x_1), s(x_2), \dots, s(x_n))$  results in an unsatisfactory estimate due to overfitting. In many situations, it simply reproduces the data. As an alternative, we define the *local likelihood* estimate of  $s(x_i)$  as

$$\hat{s}(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i} x_i \quad (11)$$

where  $\hat{\beta}_{0i}$  and  $\hat{\beta}_{1i}$  maximize the local likelihood:

$$L_i(\beta_{0i}, \beta_{1i}) = g(\{y_j, \beta_{0i} + \beta_{1i} x_j, \mathbf{c}_j\}, j \in N_i) \quad (12)$$

The local likelihood procedure produces a smooth estimate of the curve  $s(\cdot)$  at the points  $\{x_1, x_2, \dots, x_n\}$ . It avoids overfitting by averaging over neighborhoods. The width of the neighborhoods (the span) controls the smoothness of the resulting estimate—larger spans will tend to produce smoother curves.

As mentioned, the function  $L(\beta_0, \beta_1)$  need not be a likelihood, (in Cox's model it is a "partial likelihood"), but in any case, we call this procedure "Local Likelihood" estimation.

## 2.5. Local Likelihood— Definition in the i.i.d. Case

In the i.i.d case, we observe  $n$  independent data pairs  $\{(x_1, y_1), \dots, (x_n, y_n)\}$  and we assume that given  $X = x$ ,  $Y$  has density

$$Y | x \sim f(Y, \theta) \quad (13)$$

where  $\theta = s(x)$ . The likelihood is given by:

$$L(s(x_1), s(x_2), \dots, s(x_n)) = \prod_{i=1}^n f(y_i, \theta_i) \quad (14)$$

## 10 Section 2: Local Likelihood—A description

where  $\theta_j = s(x_j)$ .

The local likelihood estimate of  $s(x_i)$  is

$$\hat{s}(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i}x_i \quad (15)$$

where  $\hat{\beta}_{0i}$  and  $\hat{\beta}_{1i}$  maximize the local likelihood:

$$L_i = \prod_{j \in N_i} f(y_j, \beta_{0i} + \beta_{1i}x_j) \quad (16)$$

### 2.6. Asymptotic Properties of Local Likelihood Estimates

Other than the fact that it produces smooth estimates, why is the local likelihood procedure reasonable? On a heuristic level, it's easy to see that for well behaved  $s(\cdot)$  functions, if  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are consistent estimators in the global case (e.g. m.l.e's), then  $\hat{s}(x_i)$  will be consistent for  $s(x_i)$ . Consider a fixed point  $x_i$ . As  $n \rightarrow \infty$  and the neighborhoods shrink in such a way that  $k_n$ , the span for sample size  $n$ , goes to infinity, we have  $\hat{\beta}_{0i} \rightarrow \beta_{0i}$ ,  $\hat{\beta}_{1i} \rightarrow \beta_{1i}$ , ( $\beta_{0i}$  and  $\beta_{1i}$  being the true slope and intercept) and the error in approximation (7) goes to zero. Hence  $\hat{s}(x_i)$  will converge to  $s(x_i)$ . When  $\hat{\beta}_{0i}$  and  $\hat{\beta}_{1i}$  are local maximum likelihood estimates, the local likelihood estimates also enjoy the optimality properties of m.l.e's. They are asymptotically normal and first order efficient with respect to sample size  $k_n$ . These properties are discussed in Section 5.

### 2.7. The Bias—Variance Tradeoff

The span parameter controls the smoothness of the estimated function. Larger spans will tend to produce smoother, less variable estimates, but these estimates will tend to be biased if the underlying function is non-linear. Conversely, smaller spans will produce less biased but more variable estimates. A data-based criterion is therefore needed to select the span that best trades off these two factors for a given data set. We describe such a criterion in Section 2.12.

### 2.8. Computation of Local Likelihood Estimates in the i.i.d. Case

To find each  $\hat{\beta}_i = (\beta_{0i}, \beta_{1i})$  we use a Newton-Raphson search. Let  $U_i(\beta^0)$  be the 2 by 1 score vector with  $j$ th entry

$$U_i(\beta^0) = \left( \frac{\partial \log L_i}{\partial \beta_{ji}} \right)_{\beta=\beta^0} \quad (17)$$

and  $I_i(\beta_0)$  be the 2 by 2 observed information matrix with  $jk$ th entry

$$I_i(\beta^0) = - \left( \frac{\partial^2 \log L_i}{\partial \beta_{ji} \partial \beta_{ki}} \right)_{\beta=\beta^0} \quad (18)$$

for the  $i$ th local likelihood both evaluated at some point  $\beta^0$ . Then given an initial guess  $\hat{\beta}_i^{init}$ , the Newton-Raphson method produces the new trial value:

$$\hat{\beta}_i^{new} = \hat{\beta}_i^{init} + I(\hat{\beta}_i^{init})^{-1} U(\hat{\beta}_i^{init}) \quad (19)$$

This procedure is iterated until convergence. It is used to find  $\hat{\beta}_i$  (and hence  $\hat{s}(x_i)$ ) for each neighborhood, going in order as  $i$  runs from 1 to  $n$ . The local likelihood estimate  $\hat{\beta}_i$  is used as a starting value for the maximization of  $L_{i+1}$ ; because the estimates don't tend to differ much from one neighborhood to the next, convergence is typically achieved in 1 or 2 iterations.

Since an  $O(k_n)$  operation is required for each neighborhood, the the entire procedure is  $O(n^2)$  (assuming  $k_n \sim n$ ). This is not a problem for moderate  $n$  (say  $n \sim 200$ ) because of the small number of iterations required. For larger data sets, we speed up the procedure by calculating the fit only every  $m$ th point; this reduces the running time by about a factor of  $m$ . The smooths for the remaining  $x$ -values are obtained by interpolation.

The scatterplot smoother of Friedman and Stuetzle (1982) uses updating formula to achieve an  $O(n)$  algorithm. We have been unable to obtain such formulae for this problem because of the non-linear nature of the estimation.

## 2.9. Exponential Family Case

A special case of the above occurs when  $f$  is a member of the exponential family. Then the log likelihood has the form

$$\log L = \sum_1^n \{y_j \theta_j - b(\theta_j) - c(y_j, \sigma)\} \quad (20)$$

where  $\theta_j = s(x_j)$  and  $\sigma$  is a scale parameter. If  $\sigma$  is unknown, (20) is not generally an exponential family but the estimation procedure we will describe is unchanged because the local score function for  $\theta$  doesn't involve  $\sigma$ .

The local log likelihood is:

$$\log L_i = \sum_{j \in N_i} \{y_j \theta_{ij} - b(\theta_{ij}) - c(y_j, \sigma)\} \quad (21)$$

where  $\theta_{ij} = \beta_{0i} + \beta_{1i} x_j$ . Letting  $X$  represent the  $n$  by 2 design matrix with first column  $(1, 1, \dots, 1)^t$  and second column  $(x_1, x_2, \dots, x_n)^t$ , and letting  $W = \text{diag}\{I(j \in N_i)\}$ , the score

function has the simple form

$$U_i(\beta_i) = X'W(y - b'(X\beta)) \quad (22)$$

The observed information is  $I_i(\beta_i) = X'Wb''(X\beta_i)X$  and the Newton-Raphson step is:

$$\hat{\beta}_i^{\text{new}} = \hat{\beta}_i^{\text{init}} + I_i^{-1}(\hat{\beta}_i^{\text{init}})X'W(y - b'(X\hat{\beta}_i^{\text{init}})) \quad (23).$$

In the above, we have modelled the natural parameter  $\theta$ . We could just as well model some other parameter (like  $E(y)$ ); in any specific problem, there may be reasons to prefer one parametrization to another. For example, in the binary response problem, it is more convenient to model the natural parameter  $\log \frac{p}{1-p}$  than the expectation  $p$  because the latter would require that the estimated smooth stay between 0 and 1.

## 2.10. Relationship to Generalized Linear Models

Model (20) can be viewed as an extension of the class of *generalized linear models* (Nelder and Wedderburn (1972)). A generalized linear model is defined by  $Y | x \sim f(Y, \theta)$  and  $E(Y) = g(\beta_0 + \beta_1 x)$ , where  $f$  has the exponential form (20). If  $g$  (the “link function”) is invertible, this corresponds to  $g^{-1}(E(Y)) = \beta_0 + \beta_1 x$ . In the local likelihood set-up, we have generalized  $\beta_0 + \beta_1 x$  to  $s(x)$ .

## 2.11. Number of Parameters— “Degrees of Freedom”

In Section 6, we discuss an approximate method for determining how many independent parameters a local likelihood smooth is really fitting. Since the local likelihood estimate produces a function smoother than the data, we would expect that it uses less than  $n$  independent parameters. This is the case. Consider a running lines smoother with span  $s$ . Such a smoother is linear in that the fit  $\hat{y}$  can be written as  $P(s)y$  where  $P(s)$  is a *smoother matrix*.  $P(s)$  will depend on the set of  $x$  values observed, as well as the span. In traditional linear least squares estimation,  $P(s)$  is the hat matrix  $X(X'X)^{-1}X'$ . We show in Section 6 that for a running lines smoother with span  $s$ , the number of degrees of freedom used up is  $\text{trace}(P(s))$ . (This result and related results are also given in Cleveland (1979)). We also show that for *any* local likelihood fit (in the exponential family), with span  $s$ , the number of degrees of freedom is about  $\text{trace}(P(s))$ . Thus, although the matrix  $P(s)$  is only used in the estimation process of the Gaussian local likelihood model, (and not in the estimation of other local likelihood models), the *trace* of this matrix turns out to be the relevant quantity nonetheless. Note that this generalizes the result in linear estimation, in which  $P(s)$  is an idempotent projection matrix and hence  $\text{trace}(P(s)) = \text{rank}(P(s)) = p$ , the rank of the column space of  $X$ .

The quantity  $\text{trace}(P(s))$  turns out to be significantly less than  $n$ . In an example given in Section 6, with 100 data points and  $s = .5$ ,  $\text{trace}(P(s))$  is 3.65. Thus we are really fitting only 3.65 “parameters”.

We also note in Section 6, however, that the “number of parameters” or “degrees of freedom” should be used only as a rough rule of thumb. This is because the distribution involved is not chi-squared with  $\text{trace}(P)$  degrees of freedom, but is more spread out. If a precise estimate of the distribution is required (for a given data set), a simulation technique described in Section 6 can be used.

## 2.12. Span Selection

The estimation of a local likelihood smooth requires the choice of a span size. In scatterplot smoothing, one popular method for choosing the span size is *cross-validation*. For a number a trial spans, smooths are estimated leaving out each data point one by one. A cross-validation sum of squares is calculated and the span having the smallest value is selected. This is detailed in Friedman and Stuetzle (1981).

In the local likelihood problem, cross-validation turns out to be very expensive computationally. As an alternative, we use a form of *Akaike’s Information Criterion (AIC)*. In fitting a generalized linear model with maximized likelihood  $L$  and  $p$  independent parameters, the *AIC* is defined by:

$$AIC = -2 \log L + 2p \quad (24)$$

The first term measures the goodness of fit of the model, while the second term penalizes the number of parameters used. Hence the *AIC* attempts to tradeoff variability and bias.

We make use of the *AIC* criterion for selecting the span of the local likelihood smooth. Using  $\text{trace}(P(s))$  as an approximate number of degrees of freedom, the span size  $s$  is selected to minimize *AIC* based on the value of the global likelihood (10). In a number of examples, we’ll see that this procedure chooses reasonable span sizes, producing estimates that aren’t too jagged nor too biased. We justify this use of *AIC* in Section 6.

## 2.13. Effects of Sample Size

The local likelihood technique, being non-parametric in nature, will work best in larger samples (say  $n \geq 100$ ). Sample size is not the only factor, however—the amount of (non-linear) structure in the data relative to the noise component is also important. We’ve found that the technique can pick up non-linear structure in some data sets with as few as 60 observations. More often, however, the local likelihood algorithm chooses a large span for small data sets, and the resultant estimate closely resembles the global linear estimate.

### 2.14. Handling of Ties

For data with tied  $x$  values, two things are done. First, each neighborhood is expanded (if necessary) to ensure that if a point  $j$  is in a given neighborhood, so is any other point  $k$  having  $x_k = x_j$ . This makes the estimation procedure invariant to the incoming order of the data points. Secondly, the smooths for each of the tied values are averaged and each smooth value is assigned the average. That is, if  $x_j = x_{j+1} \dots = x_{j+m}$ , then for each  $j \leq i \leq j+m$ ,  $\hat{s}(x_i)$  is assigned the value  $\sum_j^{j+m} \hat{s}(x_i)/(m+1)$ .

### 2.15. Multiple Covariates and Backfitting

The above discussion shows how the local likelihood idea can be used to estimate the smooth for a single covariate. If  $p$  covariates are available, there are two ways the model could be generalized. One could assume  $\theta = s(\cdot, \cdot, \dots, \cdot)$ , a smooth  $p$  dimensional function, and then estimate  $s$  by local averaging in  $p$ -dimensional space. This procedure would suffer from the so-called *curse of dimensionality*: in high dimensions, averaging neighborhoods grow very large. Consider for example a data cloud uniformly distributed in a ten dimensional unit cube. Then a hypercube-shaped neighborhood containing 10% of the data points would have sides of length .8! (since  $.8^{10} \approx .1$ ). Hence the local averaging would not be “local” at all.

To avoid the curse of dimensionality, we can smooth one co-ordinate at a time. The model takes the form  $\theta = \sum_{j=1}^p s_j(\cdot)$ . This model is less general than the  $p$ -dimensional smooth model, but one can make it more general by allowing smooth functions of products, e.g.  $s(x_i x_j)$ . To estimate the  $s_j(\cdot)$ 's, a forward stepwise algorithm is used. This algorithm is analogous to a forward stepwise regression procedure. The algorithm proceeds by smoothing on each variable, and selecting the smooth that most improves the fit. When one smooth is selected, the remaining variables are smoothed and the one that most improves the fit is chosen. The process is repeated until no new variable can significantly improve the fit.

Now suppose that this procedure selects a smooth  $\hat{s}_1(\cdot)$  at the first step and a smooth  $\hat{s}_2(\cdot)$  at the second step. Then the smooth  $\hat{s}_1(\cdot)$  may not be “optimal” given that  $\hat{s}_2(\cdot)$  is in the model. Hence it is desirable to re-estimate  $\hat{s}_1(\cdot)$  to accomodate  $\hat{s}_2(\cdot)$ . Now given the adjusted estimate  $\hat{s}_1^*(\cdot)$ , we can adjust  $\hat{s}_2(\cdot)$  and so on, iterating until convergence. This process is called “backfitting” (Friedman and Stuetzle(1982)). In general, with more than two smooths, whenever a new smooth is entered into the model, the smooths already in the model are adjusted to accommodate the new smooth. Specifically, all but one of the smooths are held constant and the remaining smooth is re-estimated. This is done for each smooth in turn until the fit no longer improves by a significant amount. As an example, suppose a new smooth  $\hat{s}_{r+1}(\cdot)$  is added to a model containing smooths  $\hat{s}_1(\cdot), \dots, \hat{s}_r(\cdot)$ . Then the backfitting procedure

would consist of estimating  $s_j(\cdot)$  in the model

$$\theta = \sum_{k \neq j} \hat{\delta}_k(\cdot) + s_j(\cdot) \quad (25)$$

treating  $\sum_{k \neq j} \hat{\delta}_k(\cdot)$  as a constant. This is done for  $j$  running from 1 to  $r + 1$ . The entire cycle is repeated until convergence.

We have no proof of convergence for the backfitting algorithm, although it has converged in all the examples that we've tried. In a simple linear regression framework, with  $p$  (possibly non-orthogonal) covariates  $x_1, x_2, \dots, x_p$ , one can show that backfitting converges to the correct answer (Stuetzle (1983), personal communication). That is, if we project the current residual vector onto each covariate in turn, the residual vector converges to the correct residual vector i.e. the response minus the projection of the response onto the column space of  $x_1, x_2, \dots, x_p$ . Breiman and Friedman (1982) show convergence of a backfitting algorithm in an additive smooth model, for a restricted class of smoothers.

## 2.16. How do we select terms for the model?

This question can be addressed through examination of  $-2 \log L(\hat{y})$ , but it is customary in generalized linear modelling to work with an equivalent measure, the "deviance". The deviance is  $2 \log(L(y)/L(\hat{y}))$  which equals to  $-2 \log L(\hat{y}) + \text{constant}$ . At each stage, then, we find the smooth that decreases the deviance the most. This smooth is then added to the model if the decrease in the deviance is large compared to the number of "parameters" used up by the smooth.

An outline of the resulting algorithm is:

### Forward Stepwise Algorithm

*While (not all variables have been selected)*

*Find the smooth that decreases deviance the most*

*If decrease < threshold1 exit*

*If current model contains more than one smooth*

*Backfit smooths until decrease in deviance < threshold2*

*End While*

*Threshold1* can be set to a value determined by the number of degrees of freedom, or set to zero to allow all covariates to enter. *Threshold2* is set to some small number like .001. The output of the algorithm is  $\{\hat{s}_1(\cdot), \hat{s}_2(\cdot), \dots, \hat{s}_h(\cdot)\}$  where  $h$  is the number of smooths selected.

## 2.17. The Scale Parameter in the Exponential Family case

The exponential form (20) may or may not contain an unknown scale parameter, but in any case, the likelihood estimation procedure is unchanged because the score doesn't involve the scale. An estimate of scale is needed, however, if the deviance is to be used to assess importance of model terms. As is true for standard generalized linear models, we would fit some maximal model and use the mean deviance as our estimate of scale. This could be used to form a "scaled deviance", proceeding thereafter as if the scale were known.

In the only exponential family model we discuss (the logistic model) the scale is a function of the mean, so this issue doesn't arise. Hence we will not go into scale estimation in this paper.

## 2.18. Other Fitting Procedures

The local likelihood procedure uses local linear estimation because it works well (especially in reducing bias at the endpoints) and is simple. More sophisticated kernel-type estimates could be used to make the procedure robust and increase the smoothness of the estimated function. Borrowing ideas from scatterplot smoothing (see e.g. Cleveland (1978)), we can downweight points based both on their distance from the center of the neighborhood and the size of the residual. While we don't discuss such a procedure in general terms, a downweighting algorithm for the proportional hazards is discussed in Section 4.11.

### 3. Application to the Logistic Model.

#### 3.1. Introduction

In Section 2, we discussed how the local likelihood technique could be applied to any generalized linear model. Probably the most commonly used such model (besides the normal regression model) is the linear logistic model for binary data. In this section, we'll briefly illustrate the local likelihood procedure in this setting; full details can be found in Hastie (1983)

#### 3.2. The Problem and a Review of the Linear Logistic Model

We have data of the form  $\{(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)\}$  where the response  $y$  is 0 or 1 and  $x$  is an explanatory variable. The observations are assumed to be independent. The problem is to investigate the dependence of  $y$  on  $x$ .

Let  $\mathbf{x} = (1, x)$  and let  $p(\mathbf{x}) = P(y = 1 | \mathbf{x})$ . The log likelihood of the data is

$$\log L = \sum_{j=1}^n \{y_j \log p_j + (1 - y_j) \log(1 - p_j)\} \quad (26)$$

where  $p_j = p(\mathbf{x}_j)$ . Letting  $X$  represent the matrix with  $j$ th row equal to  $(1, x_j)$ , the score equation has the form

$$X'(\mathbf{y} - \mathbf{p}) = 0 \quad (27)$$

The *linear logistic* model assumes that

$$\text{logit } p(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} \quad (28)$$

Written as a function of  $\boldsymbol{\beta}$ , the log likelihood is

$$\log L(\boldsymbol{\beta}) = \sum_{j=1}^n \{y_j \mathbf{x}_j' \boldsymbol{\beta} - \log(1 + e^{\mathbf{x}_j' \boldsymbol{\beta}})\} \quad (29)$$

A Newton-Raphson procedure is typically used to find  $\hat{\boldsymbol{\beta}}$ . The expected information matrix is

$$I(\boldsymbol{\beta}) = X' \text{Diag}\{p_j(1 - p_j)\} X \quad (30)$$

and the Newton-Raphson iteration has the form

$$\hat{\boldsymbol{\beta}}_{\text{new}} = \hat{\boldsymbol{\beta}}_{\text{old}} + I^{-1}(\boldsymbol{\beta}_{\text{old}}) X'(\mathbf{y} - \hat{\mathbf{p}}_{\text{old}}) \quad (31)$$

**3.3. The Local Likelihood Generalization**

The formulation of subsection 2.3 for generalized linear models can be applied directly. Instead of assuming a linear form for *logit*  $p(\mathbf{x})$ , we assume

$$\text{logit } p(\mathbf{x}) = s(\mathbf{x}) \quad (32)$$

The local likelihood for  $\mathbf{x}_i$  is

$$\log L_i(\beta_i) = \sum_{j \in N_i} \{y_j \mathbf{x}_j^t \beta_i - \log(1 + e^{\mathbf{x}_j^t \beta_i})\} \quad (33)$$

Letting  $\hat{\beta}_i$  maximize  $\log L_i(\beta_i)$ , the local likelihood estimate of  $s(\mathbf{x}_i)$  is  $\hat{s}(\mathbf{x}_i) = \mathbf{x}_i^t \hat{\beta}_i$ . With multiple covariates, the model takes the form

$$\text{logit } p(\mathbf{x}) = \sum_1^p s(\cdot) \quad (34)$$

A forward stepwise algorithm is used to select covariates, as described in Section 2.

**3.4. Example 1: Breast Cancer Data**

A study conducted between 1958 and 1970 at the University of Chicago's Billings Hospital concerned the survival of patients who had undergone surgery for breast cancer (Haberman (1976)). There are 306 observations on 4 variables.

$y=1$  if patient survived  $\geq 5$  years, 0 otherwise

$x_1$ =age of patient at time of operation

$x_2$ =year of operation

$x_3$ =number of positive axillary nodes detected

The local likelihood procedure applied to all three covariates produced the smooths shown in Figures (7) , (8) , and (9) . Table 1 shows the decrease in deviance due to each variable.

**Table 1. Analysis of Breast Cancer Data**

<i>Model</i>	<i>Deviance</i>	<i>Number of Parameters</i>
Constant	353.67	1
Age(span= .6)	346.08	2.41
Age + Yr of Oper(span .5)	343.91	2.41 + 2.54
Age+Yr of oper+ # of nodes(span .5)	307.74	2.41 + 2.54 + 2.41

Age and number of nodes are important, year of operation is not. The final model has a deviance of 307.74 on  $(306-2.41-2.54-2.41)=298.54$  degrees of freedom.

Landwehr et al (1984) analyzed this data set to explore the usefulness of partial residual plots in identifying parametric forms of covariate effects. Their final model was

$$\text{logit } p(\mathbf{x}) = \beta_0 + x_1\beta_1 + x_1^2\beta_2 + x_1^3\beta_3 + x_2\beta_4 + x_1x_2\beta_5 + (\log(1 + x_3))\beta_6 \quad (35)$$

The deviance of this model is 302.3 on 299 degrees of freedom. The fitted terms for each covariate are super-imposed on Figures (7) , (8) , and (9) (broken lines). The functions for  $x_1$  and  $x_3$  are similar; they differ for  $x_2$ , but the overall effect of this variable is very small.

Hastie (1983), Hastie (1984) and Hastie, Tibshirani and Owen (1984) discuss the relative merits of the local likelihood and partial residual plot procedures. They give two reasons to suggest why the local likelihood procedure is preferable:

- The partial residual technique, in suggesting the parametric form for a covariate effect, relies on the assumption that the covariate forms for other effects are correct. Indeed these effects are usually assumed to be linear. The local likelihood procedure finds the best functional form for all covariates simultaneously.
- The partial residual technique requires quite a bit of ingenuity in identifying the various covariate effects. The local likelihood procedure, on the other hand, is automatic.

### 3.5. Comparison to the Scatterplot Smoothing Approach

The local likelihood method extends the linear logistic model through a type of local averaging within the likelihood framework. Computationally, it would seem simpler to ignore the fact that the  $y$ 's are 0's and 1's and apply scatterplot smoothing techniques directly. This works fine for a single covariate: a scatterplot smooth of  $y$  on  $x_1$  is shown in Figure (10) . On the same figure, the estimated local likelihood probability smooth  $\exp(\hat{s}(x_1))/(1 + \exp(\hat{s}(x_1)))$  is shown (broken line). Not surprisingly, the two smooths are very similar.

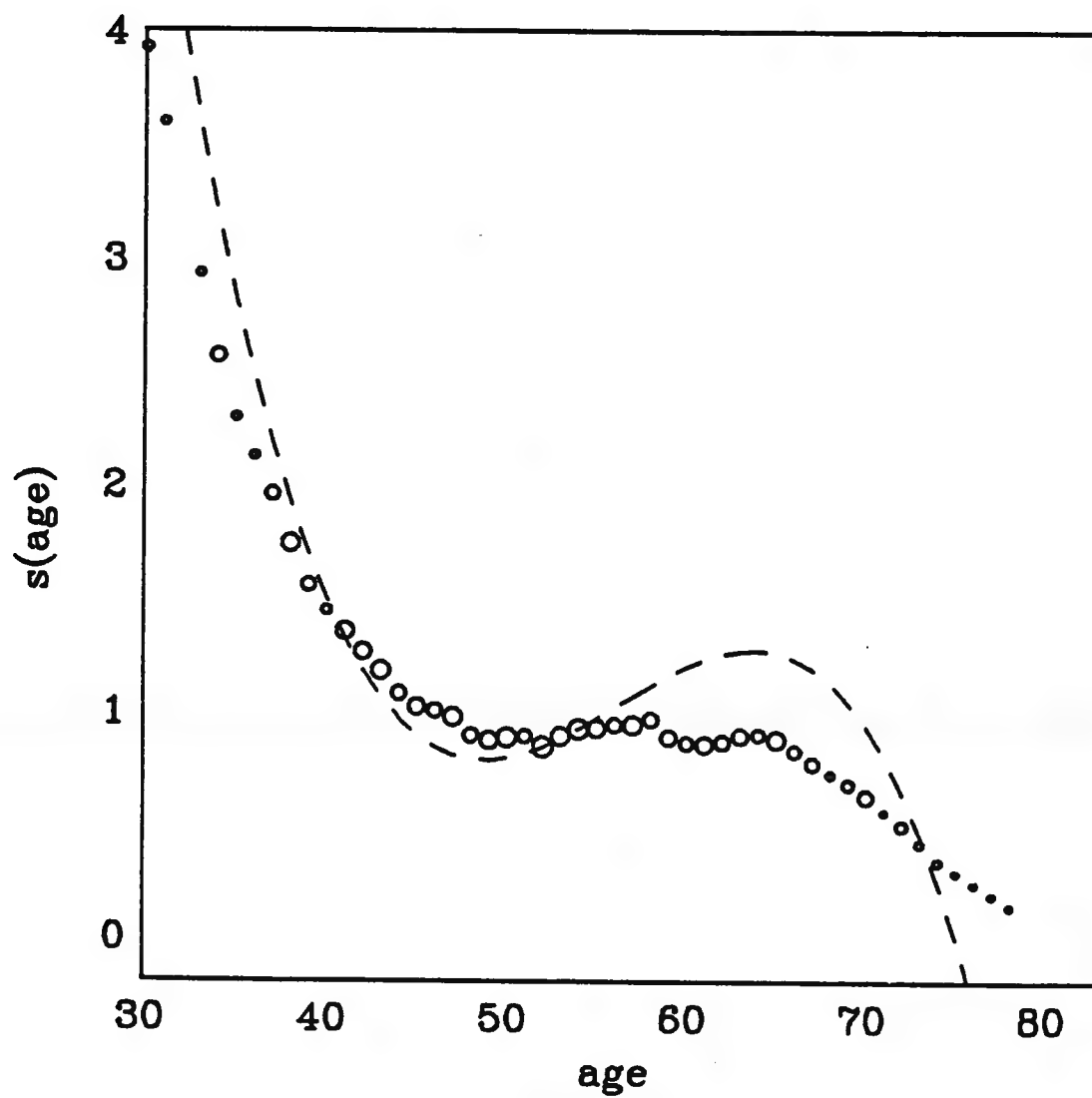
With multiple covariates, the local likelihood approach is more attractive for precisely the same reasons that the linear logistic model has gained popularity. In fitting a model of

the form  $y = \sum_1^p s(x_i)$ , one would have to ensure that the fitted probabilities lie between 0 and 1. This would require some sort of truncation of the smooths. On the other hand, the local likelihood approach models *logit p* so the fitted probabilities are always between 0 and 1. Secondly, the local likelihood approach produces an additive model on the logit scale. A large body of literature suggests that for many types of data, effects are more likely to be additive on the logit scale than on the probability scale. One could try to adapt the regression approach by grouping the  $y$ 's then using the *logit* of the grouped values as responses. This would likely produce similar results to the local likelihood approach *if* the information loss due to grouping wasn't too large.

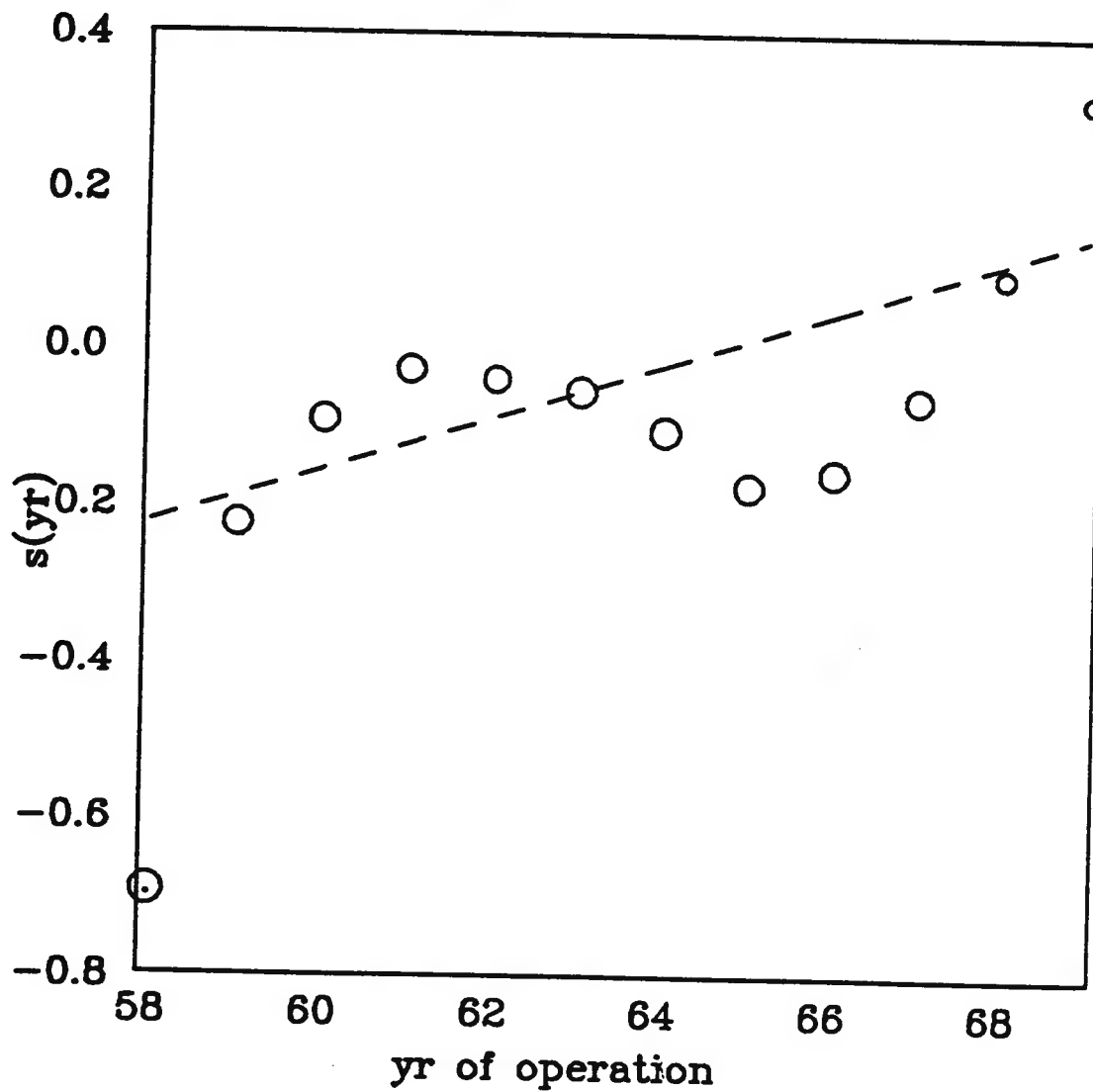
#### **A note on the figures**

Many of the figures in this section and the next section use circles to represent fitted smooth values. The area of each circle is proportional to the number of data points it represents. This is designed to indicate the density of points in each part of the smooth.

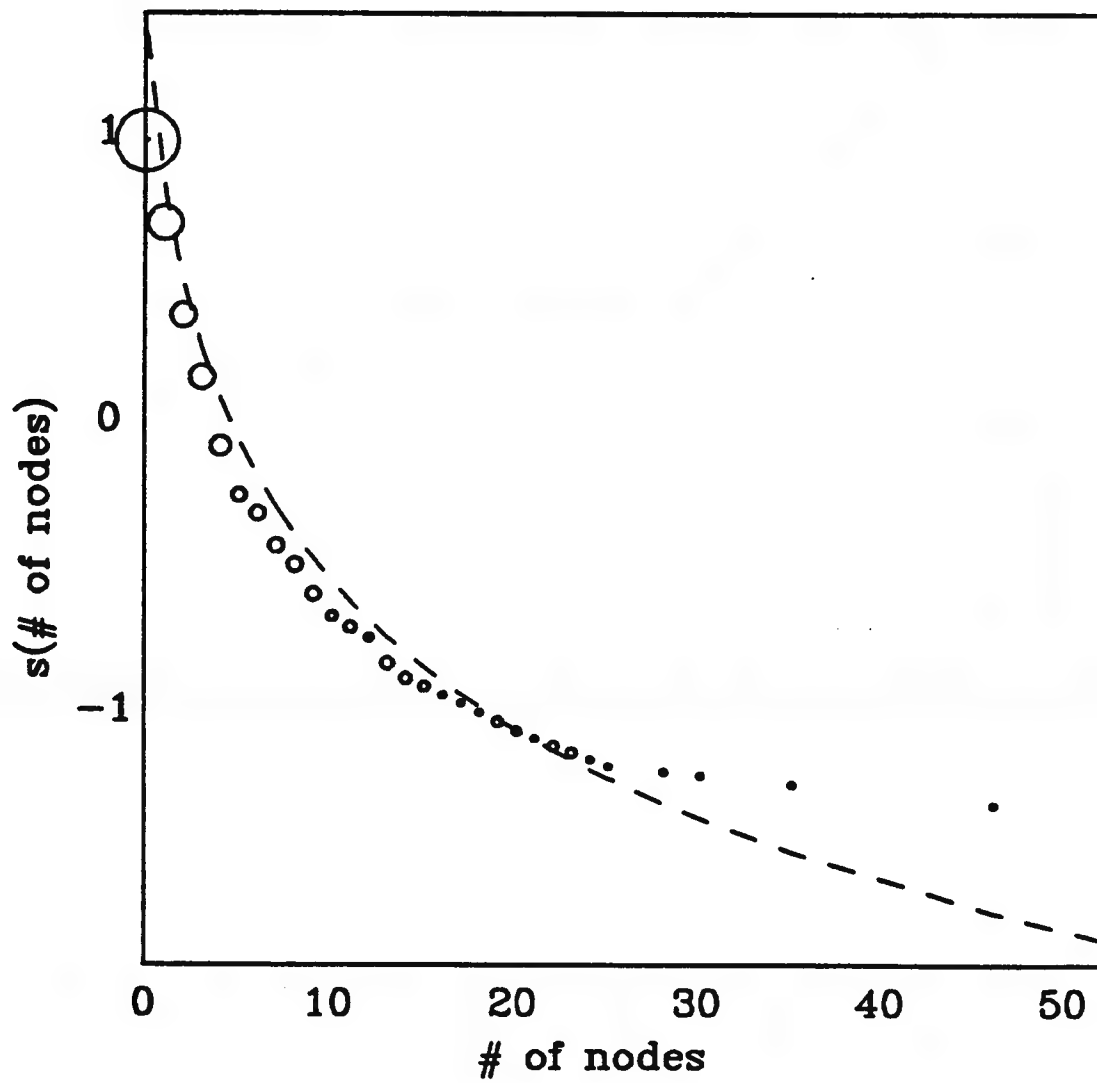
**Figure (7)**  
**Estimates for Age**  
*Circles: L.L smooth, Broken line: parametric function*



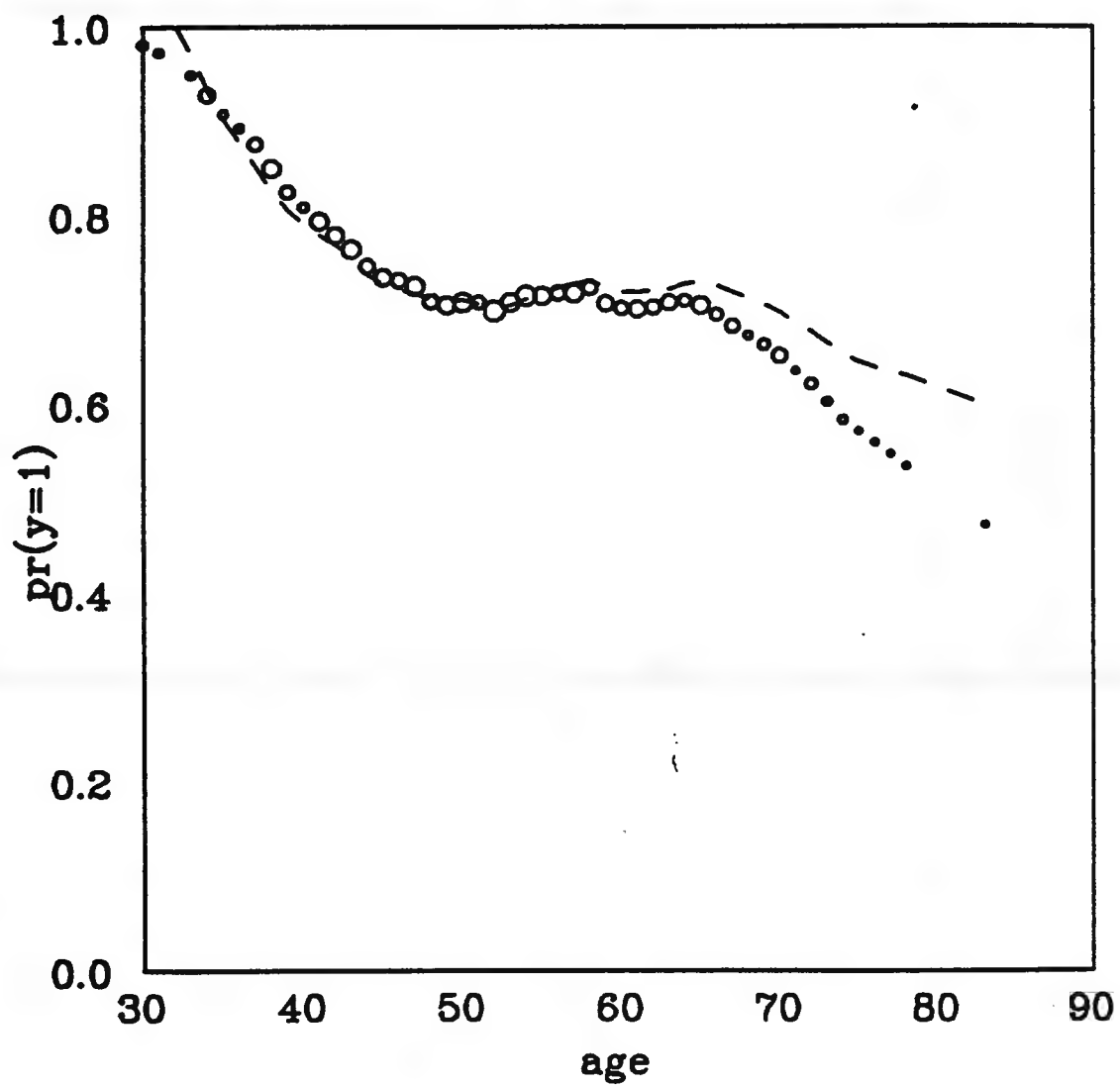
**Figure (8)**  
Estimates for Year of operation  
*Circles: L.L smooth, Broken line: parametric function*



**Figure (9)**  
Estimates for # of nodes  
*Circles: L.L smooth, Broken line: parametric function*



**Figure (10)**  
Estimates for Age  
*Circles: L.L smooth, Broken line: Scatterplot smooth*



## 4. Application to Censored Data and the Cox Model.

### 4.1. Introduction

In the censored data problem we observe data triples  $(y_i, x_i, \delta_i)$ ,  $i = 1, 2, \dots, n$  where  $\delta_i$  indicates whether or not the response  $y_i$  is censored. The data are assumed to be sorted by the covariate  $x$ , that is  $x_1 \leq x_2 \leq \dots \leq x_n$ . The proportional hazards model of Cox(1972) models the relationship between  $y$  and  $x$  by assuming that  $x$  acts on the hazard function in a multiplicative way:

$$\lambda(y | x) = \lambda_0(y) \exp(\beta x) \quad (36)$$

where  $\lambda_0(y)$  is an unspecified function, and  $\lambda(t | x)$  is the hazard function at covariate level  $x$  defined by

$$\lambda(t | x) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t, x)}{\Delta t}. \quad (37)$$

This assumption allows  $\beta$  to be estimated independently of  $\lambda_0(y)$  by maximizing the *partial likelihood*:

$$PL = \prod_{i \in D} \frac{e^{\beta x_i}}{\sum_{j \in R_i} e^{\beta x_j}} \quad (38)$$

where  $D$  is the set of indices of the uncensored  $y$ 's and  $R_i$  is the risk set prior to  $y_i$ . The local likelihood generalization of (36) is

$$\lambda(y | x) = \lambda_0(y) \exp(s(x)) \quad (39)$$

and the partial likelihood of the data is

$$PL = \prod_{i \in D} \frac{\exp(s(x_i))}{\sum_{j \in R_i} \exp(s(x_j))} \quad (40)$$

where  $R_i = \{j | y_j \geq y_i\}$ , the risk set at time  $y_i - 0$ .

To estimate  $s(x_1), s(x_2), \dots, s(x_n)$ , we apply the local likelihood technique introduced in Section 2. As before, let  $N_i$  be a symmetric neighborhood around  $x_i$ :

$$N_i = \{\max(i - \frac{k-1}{2}, 1), \dots, i-1, i, i+1, \dots, \min(i + \frac{k-1}{2}, n)\} \quad (41)$$

For  $x \in N_i$  we assume  $s(x) \approx \alpha_i + x\beta_i$ , and the local partial likelihood for the data in  $N_i$  is

$$PL_i = \prod_{l \in D \cap N_i} \frac{\exp(\alpha_i + x_l \beta_i)}{\sum_{j \in R_l \cap N_i} \exp(\alpha_i + x_j \beta_i)} \quad (42)$$

To estimate  $\alpha_i$  and  $\beta_i$ , we maximize  $PL_i$ . Note, however, that  $\alpha_i$  is not estimable from  $PL_i$  since the  $\exp(\alpha_i)$  terms cancel one another giving

$$PL_i = \prod_{l \in D \cap N_i} \frac{\exp(x_l \beta_i)}{\sum_{j \in R_l \cap N_i} \exp(x_j \beta_i)} \quad (43)$$

Let  $\hat{\beta}_i$  maximize  $L_i(\cdot)$ . Although  $\alpha_i$  (thus  $s(x_i)$ ) is not estimable locally, we can use the slope estimates  $\{\hat{\beta}_1, \dots, \hat{\beta}_n\}$  to estimate  $\{s(x_1), \dots, s(x_n)\}$ , as follows. We have  $s(x_i) = \int_c^{x_i} s'(t) dt$  and  $s'(x) = \beta_i$  for  $x \in N_i$ , hence to estimate  $s(x_i)$  we can use any estimate of  $\int_c^{x_i} s'(t) dt$  based on  $(x_1, \hat{\beta}_1), \dots, (x_n, \hat{\beta}_n)$ . Before discussing some particular integral estimators, it is important to note that the choice of  $c$  is arbitrary, reflecting the fact that  $s(x)$  is only determined up to an additive constant. Substitution of  $s(x) + c$  for  $s(x)$  in (36) doesn't change the model because the factor  $e^c$  can be absorbed into the arbitrary function  $\lambda_0(t)$ . For simplicity, then, we define  $c = x_1$ , so that  $\hat{s}(x_1) = 0$ .

To estimate  $\int_{x_1}^{x_i} s'(t) dt$ , we can use the simple rectangular rule defined by

$$\hat{s}(x_i) = \sum_1^i (x_j - x_{j-1}) * \hat{\beta}_i \quad (44)$$

for  $i > 1$  and  $\hat{s}(x_1) = 0$ . This could also be written as  $\hat{s}(x_i) = (x_i - x_{i-1}) * \hat{\beta}_i$ , so that the rectangular rule constructs the estimate  $\hat{s}(\cdot)$  by joining each line segment to the previous one, with prescribed slope  $\hat{\beta}_i$ .

For greater accuracy, we instead use the trapezoidal rule defined by

$$\hat{s}(x_i) = \sum_1^i (x_j - x_{j-1}) * \frac{(\hat{\beta}_i + \hat{\beta}_{i-1})}{2} \quad (45)$$

for  $i > 1$  and  $\hat{s}(x_1) = 0$ .

The procedure is summarized in the following algorithm:

### Local Likelihood Smoother for the Cox Model

*For*  $i=1$  *to*  $n$

Find  $\hat{\beta}_i$  that maximizes  $PL_i(\cdot)$

*End For*

$\hat{s}(x_1) = 0$

*For*  $i=2$  *to*  $n$

$\hat{s}(x_i) = \sum_1^i (x_j - x_{j-1}) * \frac{(\hat{\beta}_i + \hat{\beta}_{i-1})}{2}$

*End For*

*Output*  $\{\hat{s}(x_1), \hat{s}(x_2), \dots, \hat{s}(x_n)\}$

## 4.2. Significance of a Smooth and “Degrees of Freedom”

In proportional hazard modelling, the “deviance” has no obvious analogue, so one works directly with  $-2 \log PL$  to assess significance of a smooth. The “degrees of freedom” of the smooth are more difficult to obtain, however. The simulation study in section 6 shows that the formula  $trace(P)$  is biased downward for the Cox model. Therefore, we find the mean deviance decrease by the simulation technique described in that section. The *trace* formula is still adequate for span selection, however, since biases will tend to cancel out in comparing two spans.

## 4.3. Handling Tied Survival Times

For data with tied  $t_i$  values, we use the approximation suggested by Peto(1972) and Breslow(1974) for the partial likelihood

$$PL \approx \prod_i \frac{\exp(z_i \cdot \beta)}{(\sum_{j \in R(t_i)} \exp(x_j \cdot \beta))^{d_i}} \quad (46)$$

where  $d_i$  equals the number of failures at  $t_i$  and  $z_i$  equals the sum of  $x_i$ ’s for items failing at  $t_i$ . This approximation is used for each of the partial likelihoods  $PL_i(\cdot)$ .

**4.4. Multiple Covariates**

With more than one covariate, the model takes the form

$$\lambda(t | \mathbf{z}) = \lambda_0(t) \exp\left(\sum_{j=1}^p s_j(\cdot)\right) \quad (47)$$

The smooths are estimated in a forward stepwise manner, with backfitting, as discussed in Section 2.

**4.5. Example 2: The Stanford Heart Transplant Data**

The first example that we will use for illustration of this technique is the Stanford Heart Transplant Data, as reported by Miller and Halpern(1982). There are 157 observations consisting of survival time after transplant and two covariates: age (in years) at time of transplant and T5 mismatch score. The procedure chose a span size of .7 and produced the smooth shown in Figure (11) . The actual estimate of relative risk ( $\exp(\hat{s}(\cdot))$ ) is shown in Figure (12) . A summary of the results is shown in Table 2.

**Table 2. Stanford Heart Transplant Data**  
*Analysis of Age*

<i>Model</i>	<i>-2Log Likelihood</i>	<i>Number of Parameters</i>
Null	902.40	0
Age (linear)	894.82	1
Age + Age <sup>2</sup>	886.24	2
Age (smooth, span .7)	884.65	2.95
Piecewise linear	885.40	2

The smooth reduced  $-2\log PL$  from a null value of 902.40 to 884.65. For comparison, a standard proportional hazards model with a single term for age produced a value of 894.82 for  $-2\log PL$  and the addition of a quadratic term for age reduced it to 886.24. The resulting quadratic function is shown in Figure (11) (broken line). The smooth in Figure (11) suggests that the relative risk before age 45 is approximately constant, while the quadratic curve, perhaps misleadingly, indicates a decrease in risk before age 45. We note that the smooth produces a smaller value of  $-2\log PL$  (by 1.6) but uses .95 more “parameters”.

Based on Figure (11) , we tried to summarize  $\hat{s}(\cdot)$  by a piecewise linear covariate  $z = -.2$  for  $age < 44$  and  $z = .12 \times age - 5.5$  for  $age > 44$ . Using  $z$  as a covariate in a model of the form  $\lambda_0(t) \exp(\beta \hat{s}(x))$ , a standard computer program for fitting proportional hazards models

produced a value of 885.40 for  $-2 \log PL$ . This provides further evidence that the quadratic shape for the relative risk may not be realistic.

#### 4.6. Stanford Heart Transplant Data: Age and T5

The forward stepwise algorithm was run on the Stanford Heart Transplant data described in Example 2. The smooths for each variable separately are shown in Figures (11) and (13). The threshold value was set to zero to allow both variables to enter. The results are summarized in Table 3.

**Table 3. Stanford Heart Transplant Data**  
*Analysis of Age and T5*

<i>Model</i>	<i>-2Log Likelihood</i>	<i>Number of Parameters</i>
Null	902.40	0
T5 (smooth, span= .7)	899.99	2.68
Age + T5	882.53	2.95 + 2.68
Age + T5 (backfit)	882.52	2.95 + 2.68

Age was entered first, then T5 mismatch score. The smooth for T5 is shown in Figure (14). Backfitting had only a negligible effect, so the smooth for age was virtually identical to Figure (11). The results indicate that the effect of T5, after adjusting for age, is very slight.

#### 4.7. Example 3: Mouse Leukemia Data

Kalbfleisch and Prentice (1980) analyzed the results of a study designed to examine the genetic and viral factors that may influence the development of spontaneous leukemia in AKR mice. The original data set contains 204 observations, with six covariates and 2 causes of death (cancerous and non-cancerous) measured. Kalbfleisch and Prentice perform a number of analyses; we will follow one of them here, using any death as the endpoint and the four covariates:

- $x_1$ : antibody level (%)
- $x_2$ : Gpd-1 phenotype
- $x_3$ : sex (1=male, 2=female)
- $x_4$ : coat colour

Antibody level took on continuous values, although about half of the mice had a value of 0. The other three covariates were binary. Of the 204 observations, 4 had missing values and were discarded.

Table 4 shows the results of forward stepwise local likelihood estimation applied to these data.

**Table 4. Mouse Leukemia Data**  
*Multivariate Analysis*

<i>Model</i>	<i>-2Log Likelihood</i>	<i>Number of Parameters</i>
Null	1189.06	0
Antibody (smooth, span= .5)	1173.98	1.85
Antibody+Gpd-1	1170.90	1.85 + 1
Antibody (linear)	1183.16	1
Antibody (linear + quadratic)	1183.07	2
Piecewise linear	1177.34	2

Each of GPD-1, sex and coat color were modelled with a single parameter. Antibody was the most important factor, reducing  $-2\log PL$  by 15.08. Gpd-1 was next in importance but not significant at 95%. A graph of the estimated smooth for antibody is shown in Figure (15) (the smooth values were not joined so that the distribution of antibody levels could be seen). It is markedly non-linear, changing slope at antibody level = 7.5%. Also included in Table 4 are linear and quadratic terms for antibody. Even with a quadratic term, the fit of the parametric Cox model is significantly worse than the local likelihood smooth.

Based on Figure (15), a piecewise linear covariate was created by joining each of the left and rightmost smooth values to the bending point by straight lines.  $-2\log PL$  for this covariate was 1177.34, still significantly worse than the smooth model. This indicates that the bowed shape of the smooth between antibody levels 7.5% and 80% is supported by the data.

#### 4.8. Bootstrapping the models

To assess the variability of an estimated relative risk curve, the bootstrap (Efron (1979)) can be applied. As in the regression modelling, there are (at least) two ways to bootstrap: we can resample the triples  $(y_i, x_i, \delta_i)$  or we can resample the residuals  $(r_i, \delta_i)$  (where  $r_i = \hat{\Lambda}_0(y_i) \exp(\hat{\beta}(x_i))$ ) and add them back to the fitted model. As in the regression case, the second method assumes that the fitted model is correct.

The results for these two bootstrap methods applied to Example 2 are shown in Figures (16) and (17). 20 bootstraps were computed for each method. The two plots indicate about

the same amount of variability for the low and medium age groups, but somewhat surprisingly, the residual sampling shows greater variability in the higher ages. The use of the bootstrap for the proportional hazard model requires further study; Efron(1980) looks at the bootstrap for the Kaplan-Meier curve.

#### 4.9. Case Control Data and a Comparison to Thomas' Method

Thomas (1983) provides a method of finding the maximum likelihood estimate of  $r(x)$  in the proportional hazards model  $\lambda(t | x) = \lambda_0(t)r(x)$  subject to  $\hat{r}(x)$  monotone in  $x$ . The algorithm is extremely complex and not fully understood by this author. It produces a step function  $\hat{r}(\cdot)$  with steps occurring only at some of the failures.

Thomas applied his algorithm to a data set consisting of 215 lung cancer cases, each matched with 5 controls, sampled from a large cohort of Quebec chrysotile miners and millers. The covariate of interest was total dust exposure. The effect of various levels of dust exposure was desired so that industry standards could be established.

In order to handle case control data of this type, only a small change is required in the local likelihood procedure. The local partial likelihood simply becomes a partial likelihood for case-control data. This, in turn, is the same as the partial likelihood for prospective data, except that each risk set consists of a case and its associated controls (see Prentice and Breslow (1978) for details). It turns out that in the modified local likelihood procedure, a case-control set only enters into the partial likelihood for a given neighborhood if the case and at least one control exist in the neighborhood.

Figure (18) shows the results of the various estimation procedures applied to the lung cancer data.\* The solid line is the local likelihood smooth  $\exp(\hat{s}(\cdot))$ , and the step function (dashed line) is Thomas monotone m.l.e. The functions are in qualitative agreement, with the monotone m.l.e suffering from its jagged shape.

The advantages of the local likelihood procedure over Thomas' method are clear. The monotone m.l.e is not smooth and is forced to be monotone. As well, Thomas' procedure can handle only one covariate. The local likelihood procedure suffers from none of these problems.

---

\* Unfortunately, we could only obtain a slightly smaller data set from Thomas, consisting of 188 of the 215 case-control groups. The local likelihood procedure was applied to this reduced data set, while Thomas' procedure was applied to the full data set

#### 4.10. A Bias Study

In this section we discuss a few simulations designed to investigate how well the procedure estimates the true underlying function. In particular, we want to find out how much it underestimates curvature for larger spans, especially at the endpoints.

A sample of 200  $X$  values were generated from  $U(-1, 1)$ , and survival times  $T$  were generated from the model  $\log T = 5 + 4x^2 + \epsilon$  where  $\epsilon$  had the extreme value distribution  $\exp(\epsilon - \exp(\epsilon))$ . This corresponds to the hazard model  $\lambda(t | x) = \exp(-5 - 4x^2)$ . Censoring times  $C$  were then generated from  $U(0, 11)$ , and the observed response was  $Y = \min(T, C)$ . This resulted in an average censoring rate of 51 percent. Our aim here was to found out how well the procedure reproduces curvature in the middle of the covariate range (so that endpoint effects don't enter in). Figure (19) shows the average of 20 replications (with the same set of  $x$  values) allowing the procedure to choose the span by the *AIC* criterion. Since the functions are determined only up to an additive constant, they were translated to have the same mean over the range of  $x$ . The average smooth captures the shape of the true function remarkably well.

Next, we investigated the effect of endpoint bias. We generated data from the same model as above, except that  $X$  was  $U(-1, .5)$  (We cut off the  $X$  range so that the true function would be non-linear near an endpoint.) Figure (20) shows the average of 20 replications, allowing the procedure to choose the span. The average smooth underestimates the curvature, but reproduces the function quite well.

We conclude from this modest study that the procedure has low bias, with a tendency to underestimate curvature slightly at the endpoints.

#### 4.11. A Robust Fit

There are two types of influential points that can create problems in regression modelling: outliers in time space and outliers in covariate space. The first type are not as much of a problem here because the partial likelihood depends only on the ranks of the survival times. Still, Cain and Lange (1983) give an example in which a few large survival times have a large effect on the regression coefficient.

Outliers in covariate space are potentially more dangerous. Because of the local nature of the fitting, it will not be as much a problem in the local likelihood model as it is in the linear proportional hazards model, but with spans as large as  $.7n$ , it is still a concern.

A simple modification of the fitting procedure can help reduce the effect of covariate outliers in both the standard and local likelihood proportional hazard models. The idea is to downweight observations based on their distance from the "center" of the data. This idea is exploited in the bounded influence regression literature (see Krasker and Welch (1973)

and the references therein). In order to define a “weighted” partial likelihood estimate, we need to define the partial likelihood for a sample with weights  $w_i$  on  $(y_i, x_i, \delta_i)$ ,  $i = 1, 2, \dots, n$ . ( $\sum_1^n w_i = n$ ). Consider first the case where there are no ties in the original data. It is natural to require that when the  $w_i$ ’s are integers, the weighted partial likelihood should exactly coincide with the partial likelihood for a sample with  $w_i$  copies of point  $i$ . A form suggested by Cain and Lange almost satisfies this requirement:

$$PL^w = \prod_{i \in D} \left( \frac{\exp(x_i \beta)}{\sum_{j \in R_i} w_j \exp(x_j \beta)} \right)^{w_i} \quad (48)$$

When the weights are integers, this reduces not to the exact partial likelihood for the corresponding sample, but to the standard approximation for tied data given in Section 3.2. As long as each  $w_i$  is small compared to  $\sum_{j \in R_i} w_j$ , (as it will be in our case) this approximation is adequate. When the original data contains ties, we can modify (48) :

$$PL^w = \prod_{i \in D} \frac{\exp(\sum_1^{d_i} x_j w_j \beta)}{[\sum_{j \in R_i} w_j \exp(x_j \beta)] \sum_1^{d_i} w_i} \quad (49)$$

where  $d_i$  is the number of failures at  $y_i$ . Expression (49) reduces to the correct (approximate) partial likelihood when the weights are integers.

Maximization of (49) with appropriate weights provides a more robust fitting procedure. Let  $\mathbf{x}^c$  be some “center” of the covariate space and let  $v_j$  be some scaled measure of distance of  $\mathbf{x}_j$  from  $\mathbf{x}^c$ . Then a reasonable choice of weights is  $w_j \sim e^{-v_j}$ . For the linear proportional hazards model, it would be natural to choose  $\mathbf{x}^c = \bar{\mathbf{x}}$  and

$$v_j = \mathbf{x}_j^t (X^t X)^{-1} \mathbf{x}_j \quad (50)$$

In the univariate case, this reduces to

$$v_j = \frac{(x_j - \bar{x})^2}{\sum_1^n (x_j - \bar{x})^2} + \frac{1}{n} \quad (51)$$

For the local likelihood extension of the model, we can use partial likelihood form (49) in each neighborhood, and weights proportional to  $e^{-v_j}$  where

$$v_j = \frac{(x_j - x_i)^2}{\sum_1^n (x_j - x_i)^2} + \frac{1}{k_n} \quad (52)$$

Note that  $x_i$  is used as the center of the neighborhood instead of the mean— this ensures that points near the ends receive large weights in their own neighborhoods.

Figure (21) shows the robust version of the local likelihood procedure applied to age variable (solid line). The smooth looks very similar to the unweighted smooth (broken line); this is not surprising since there are no outlying ages in the sample. Figure (22) shows the unweighted smooth (broken line) applied to the sample after having moved a failure at the

highest age (62) to 92 (only the portion of the the smooth from ages 12 to 62 is shown). The weighted smooth (dotted line) looks much like the weighted smooth applied to the original data (solid line, same as solid line in Figure (21) ). The downweighting has successfully reduced the effect of the outlying point on the overall smooth. Of course, the weighting scheme described here could be applied within the parametric setting, but we haven't pursued this.

The "robustifying" scheme discussed here is important if the local partial likelihood procedure is to be used in "auto-pilot" mode; alternatively, since each covariate is fit separately, a simple scatter diagram of each covariate should reveal any large outliers in covariate space.

In the theoretical investigations of the following sections, we'll restrict attention (for simplicity) to unweighted smoothing procedures.

#### 4.12. Extending the Model

There are (at least) two ways that the model could be extended. The first way would be to allow time-dependent covariates. In principle, this would be straightforward; as in the standard proportional hazards model, one would simply insert the "current" covariate values when constructing each term of the partial likelihood. There may be computational problems with this, however. With fixed covariates, the risk sets can be computed by "stripping off" each failure or censoring as they occur. With time-dependent covariates, however, the risk sets must be recomputed for each failure time. This would increase the cost by about a factor of  $n$ . We haven't tried implementing time-dependent covariates; this may be pursued in subsequent research.

Another way to generalize the model is to allow linear combinations of covariates to enter into the model. The form of the model would be

$$\lambda(t | \mathbf{x}) = \lambda_0(t) \exp\left(\sum (s(\boldsymbol{\alpha}_i \cdot \mathbf{x}))\right) \quad (53)$$

The vectors  $\boldsymbol{\alpha}_i$  could be found by a numerical search. This is the "Projection Pursuit Regression" idea introduced by Friedman and Stuetzle(1981). Besides the obvious computational cost, this model would suffer from a lack of interpretability.

**Figure (11)**  
Local Likelihood Estimate for Age  
*Circles: L.L smooth, Broken line: quadratic fit*

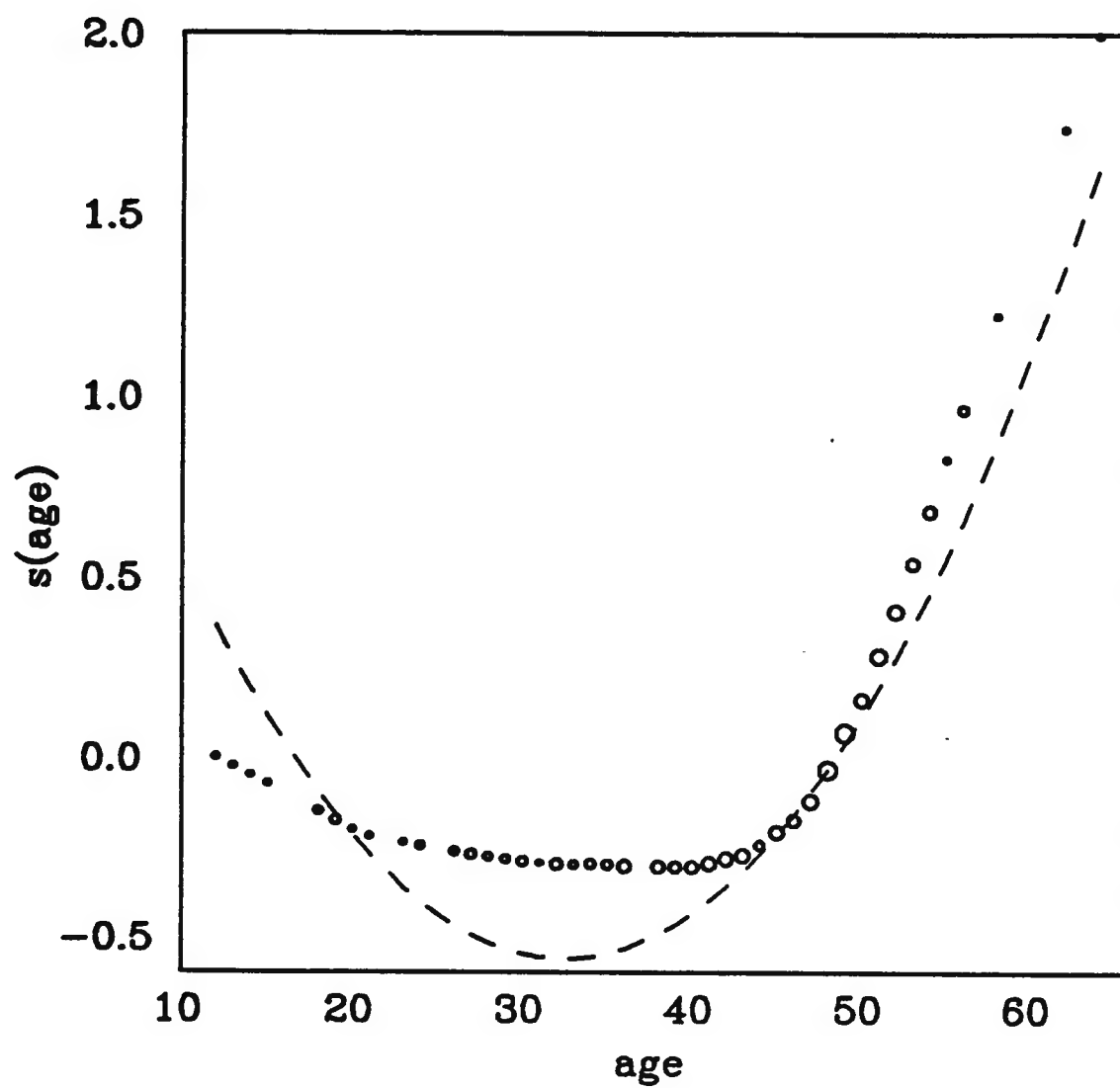
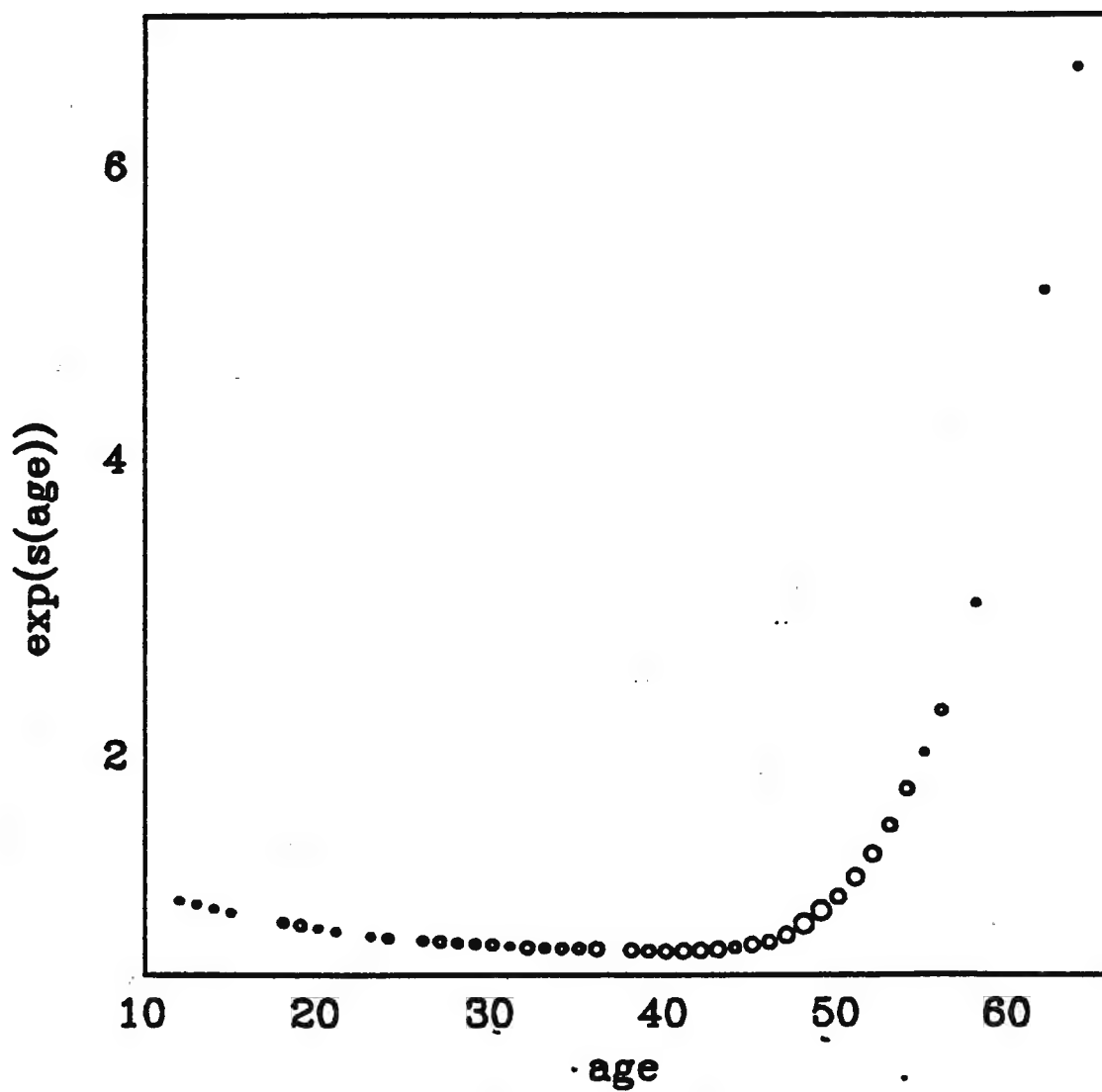
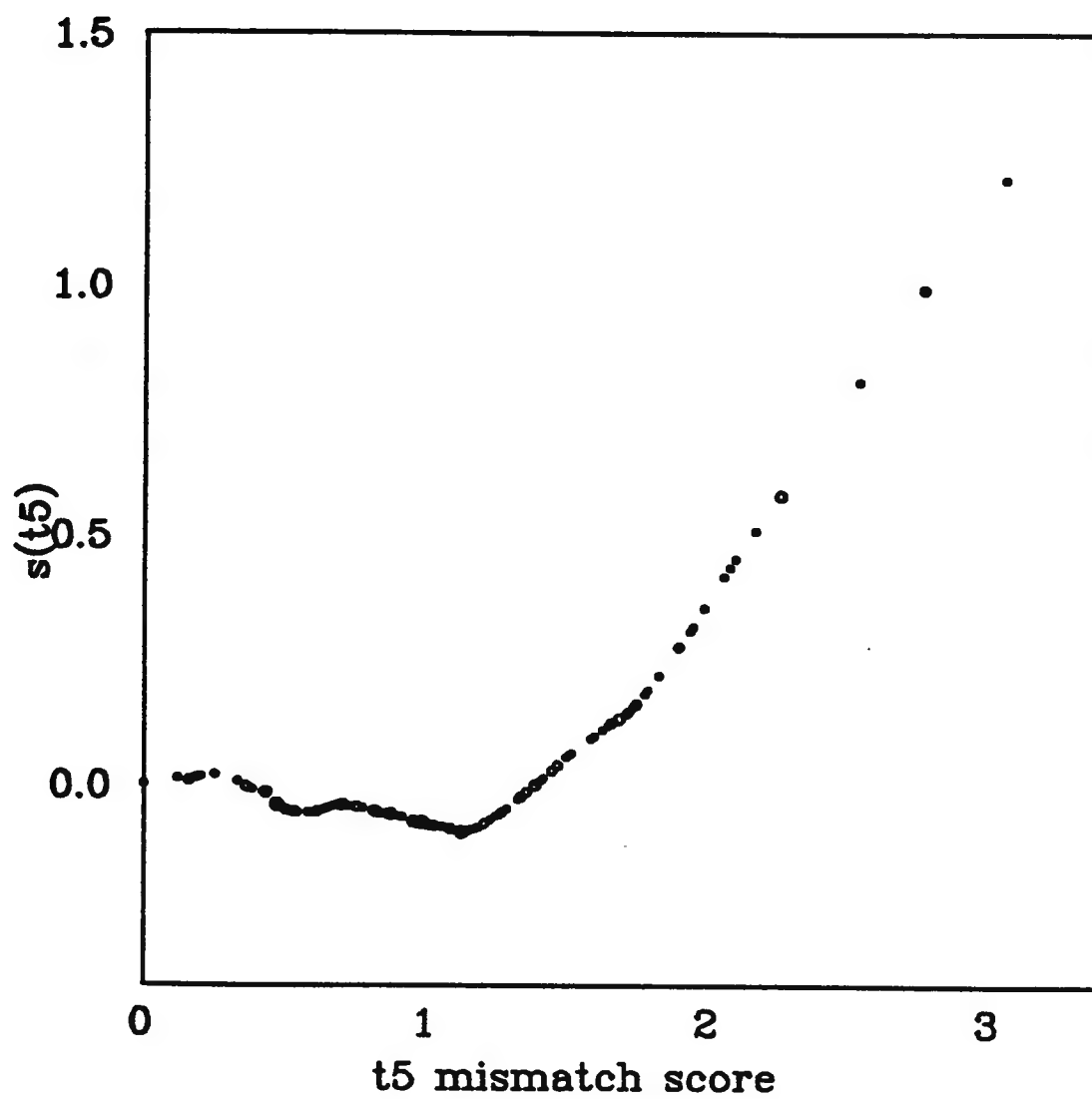


Figure (12)  
Local Likelihood Estimate of Relative risk for Age



**Figure (13)**  
Local Likelihood Smooth for T5 Mismatch Score



**Figure (14)**  
T5 smooth with age in the model

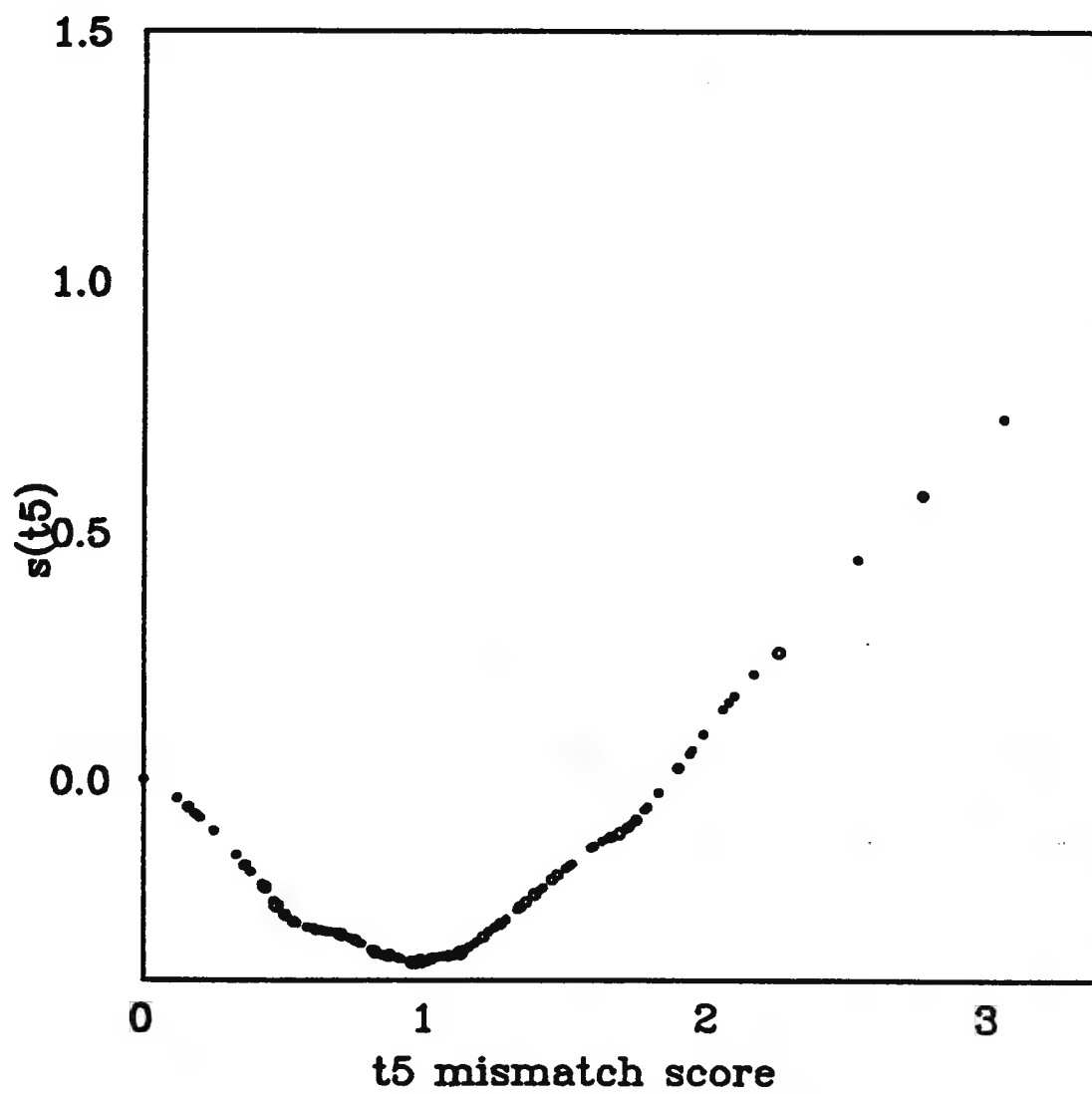
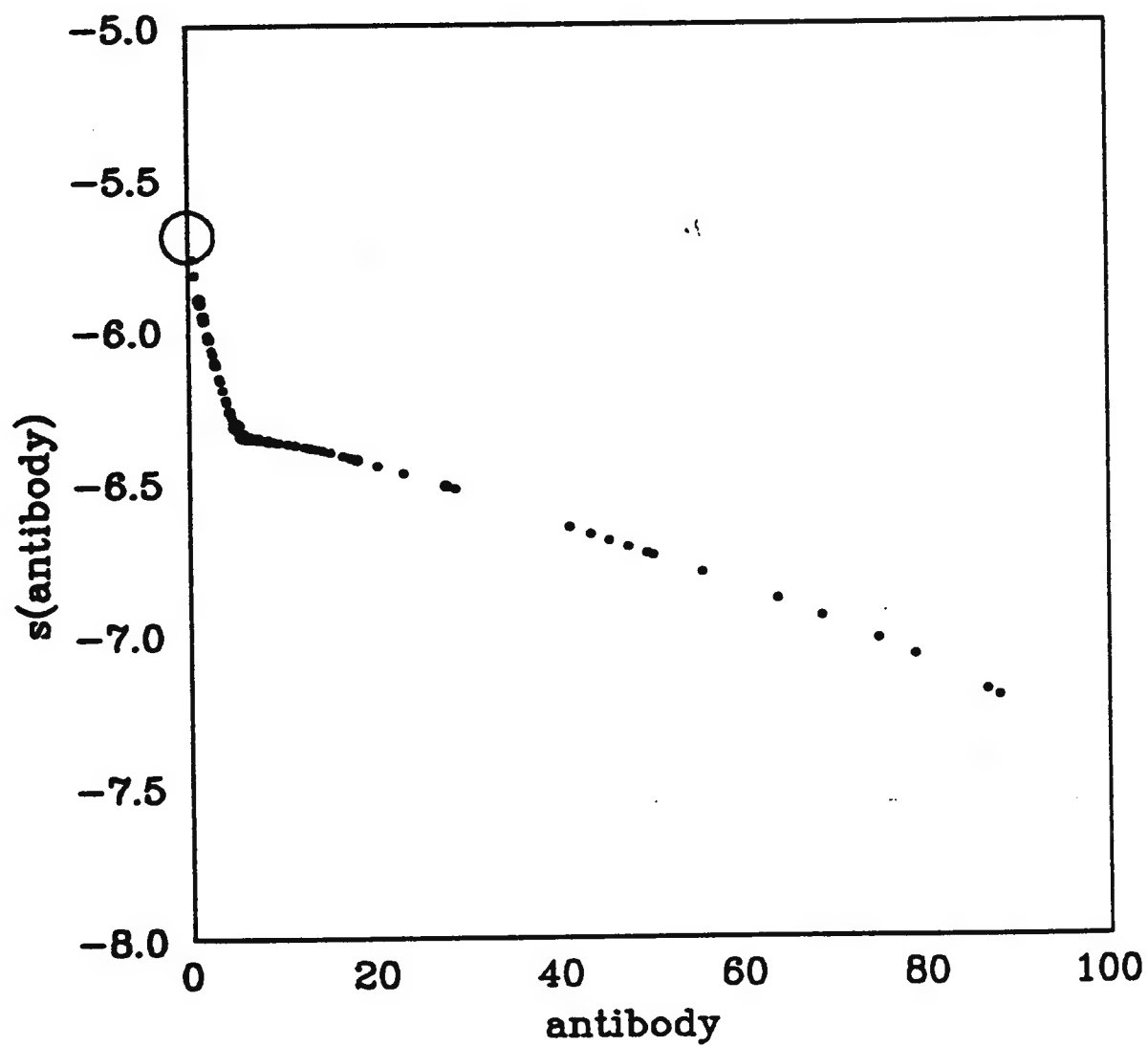
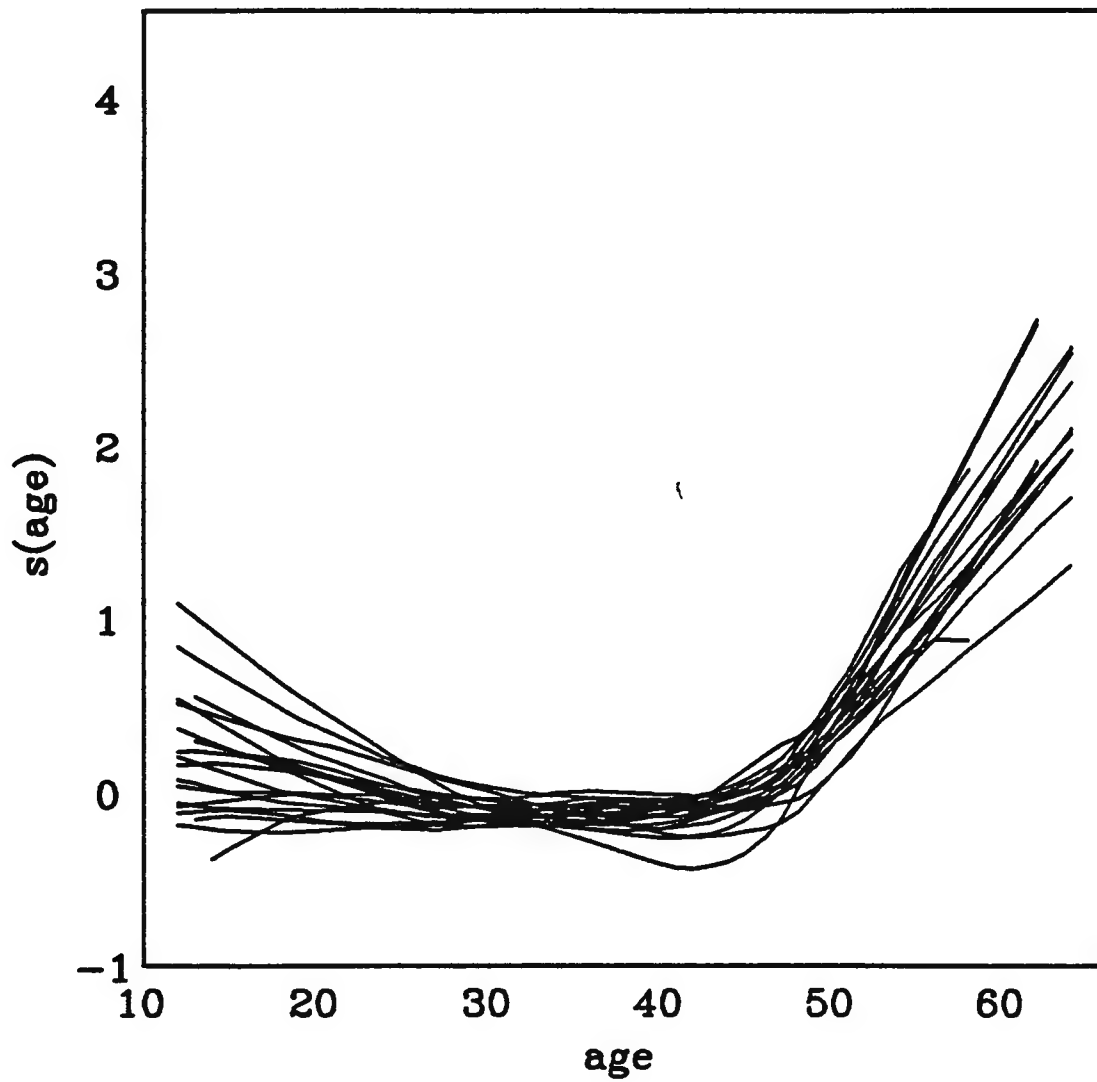


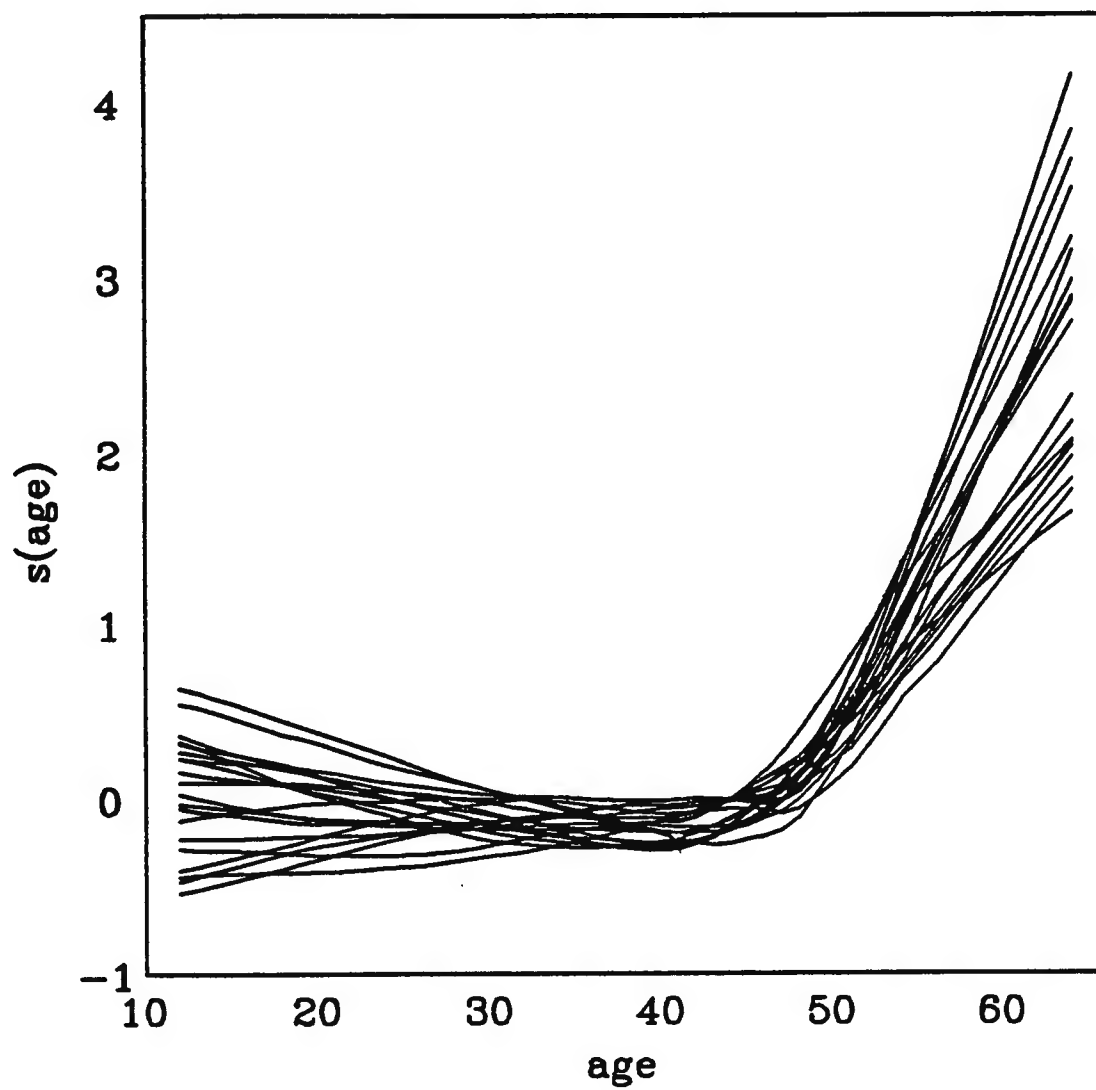
Figure (15)  
Mouse Leukemia Data: Smooth for Antibody



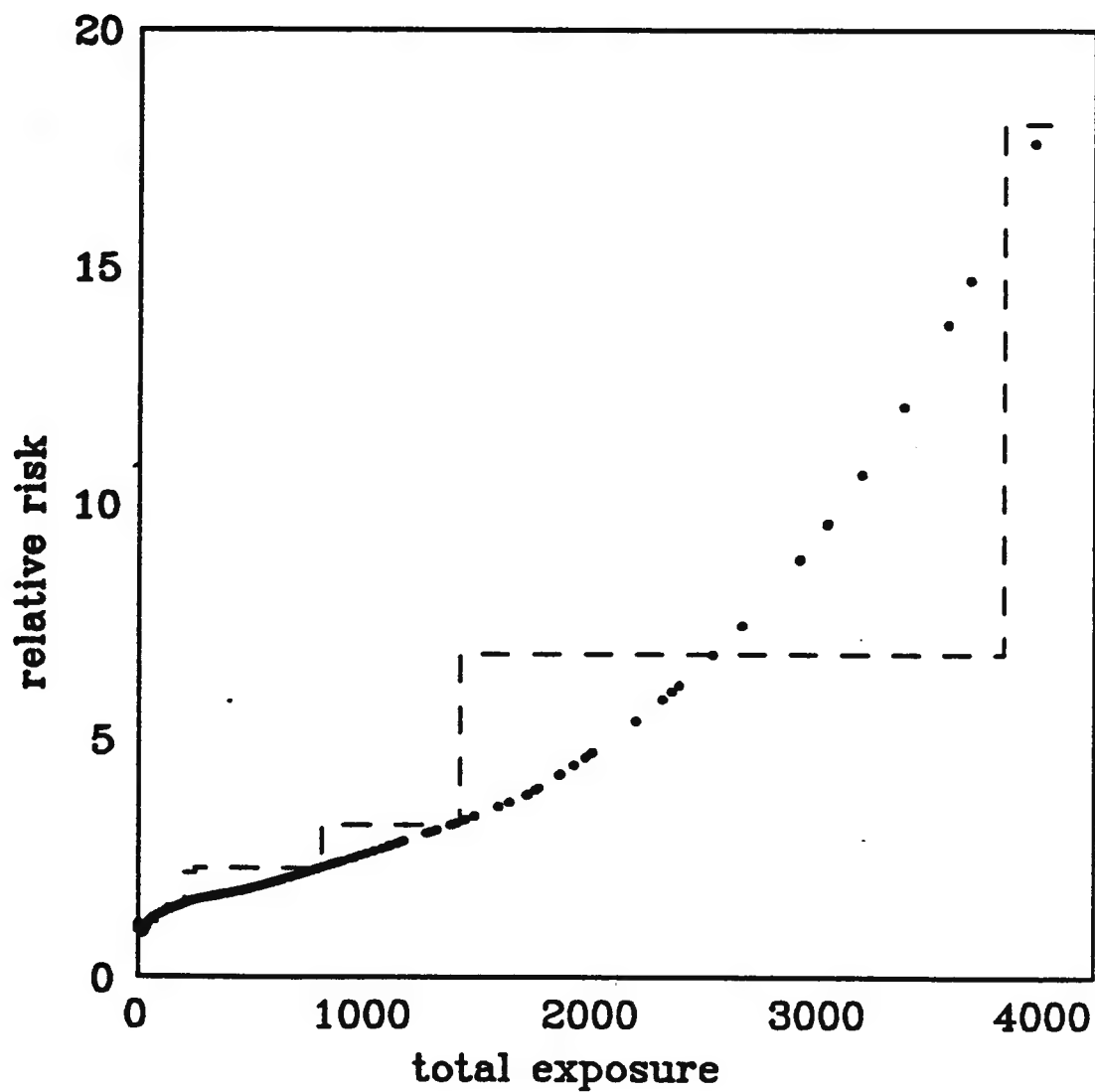
**Figure (16)**  
Bootstrap smooths (Resampling  $(y_i, x_i, \delta_i)$ )



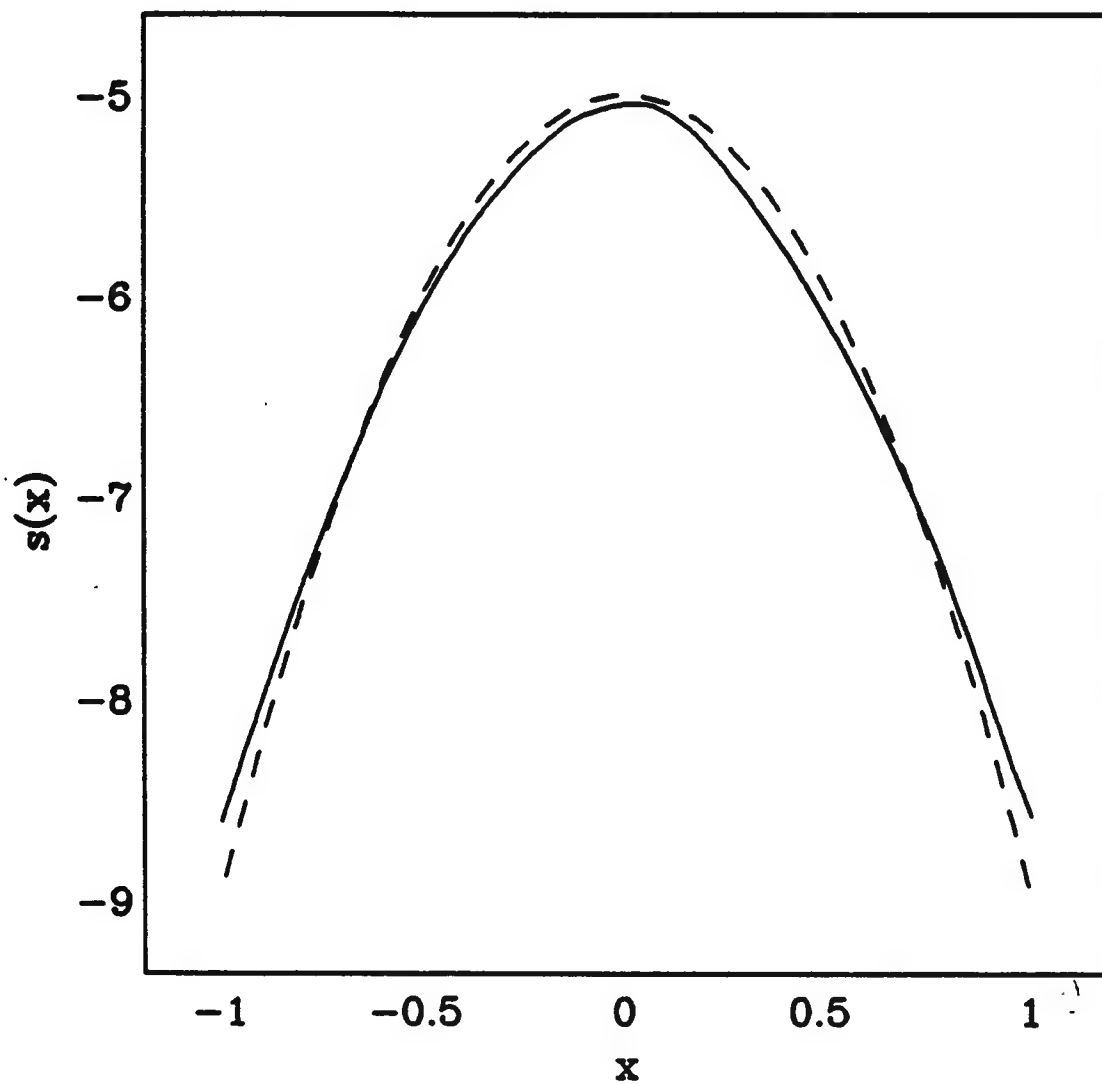
**Figure (17)**  
Bootstrap smooths (Resampling residuals)



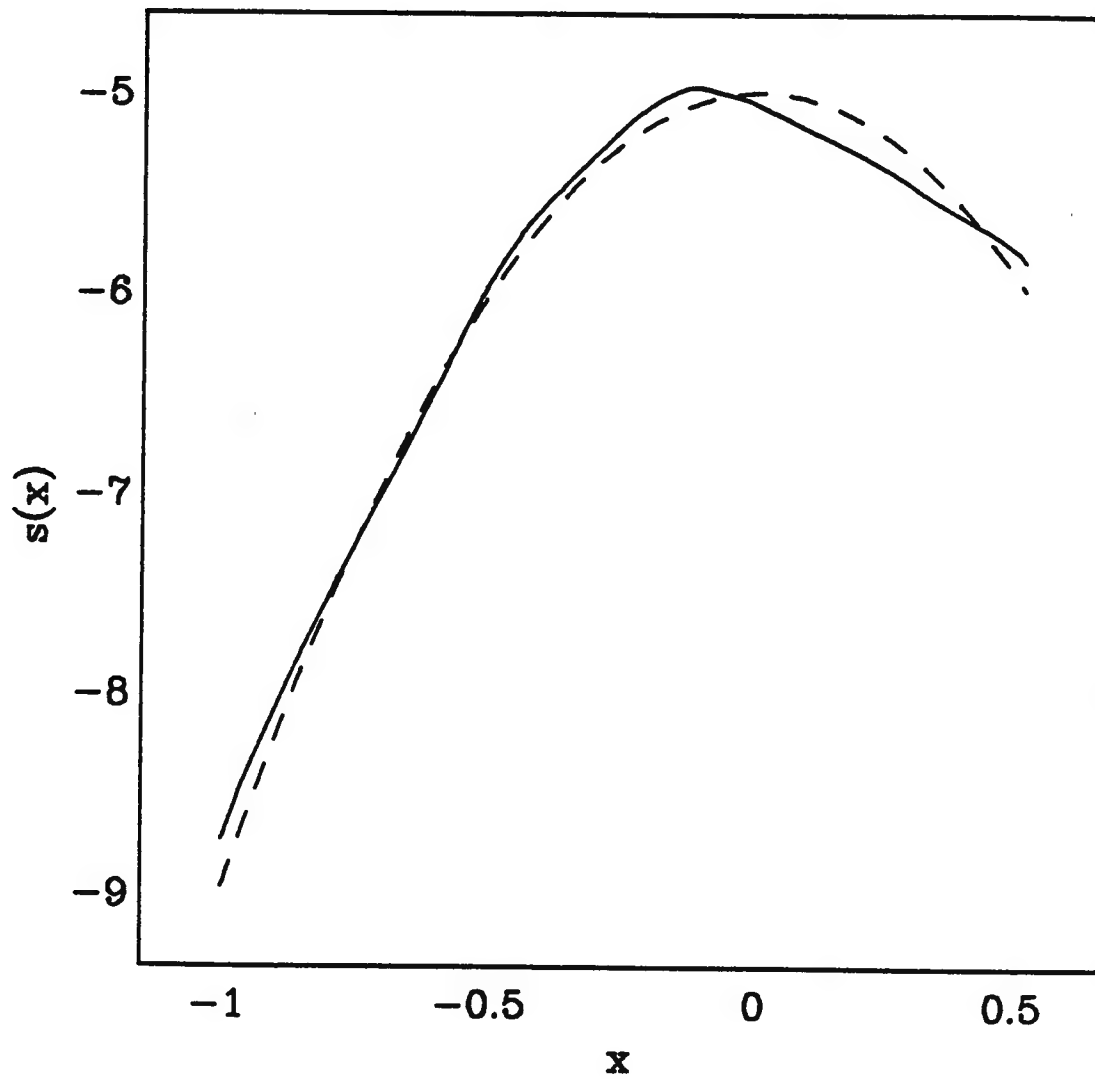
**Figure (18)**  
Estimates of Relative Risk for Lung Cancer Data  
*Circles: L.L smooth, Broken line: Monotone m.l.e*



**Figure (19)**  
Average of 20 Local likelihood fits, varying span  
*Solid line: L.L fit, Broken line: true quadratic function*

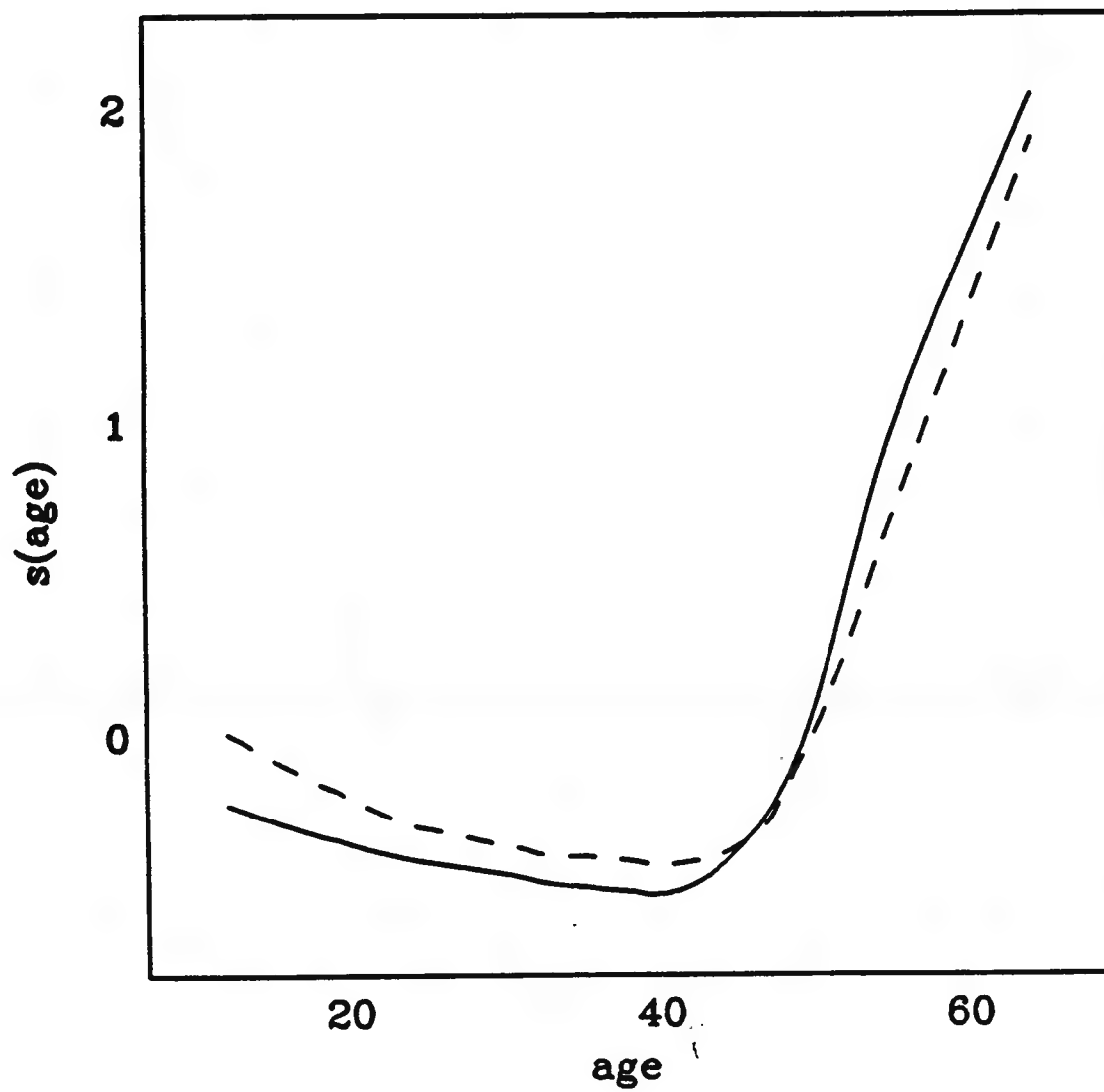


**Figure (20)**  
Average of 20 Local likelihood fits, varying span  
*Solid line: L.L fit, Broken line: true quadratic function*



**Figure (21)**

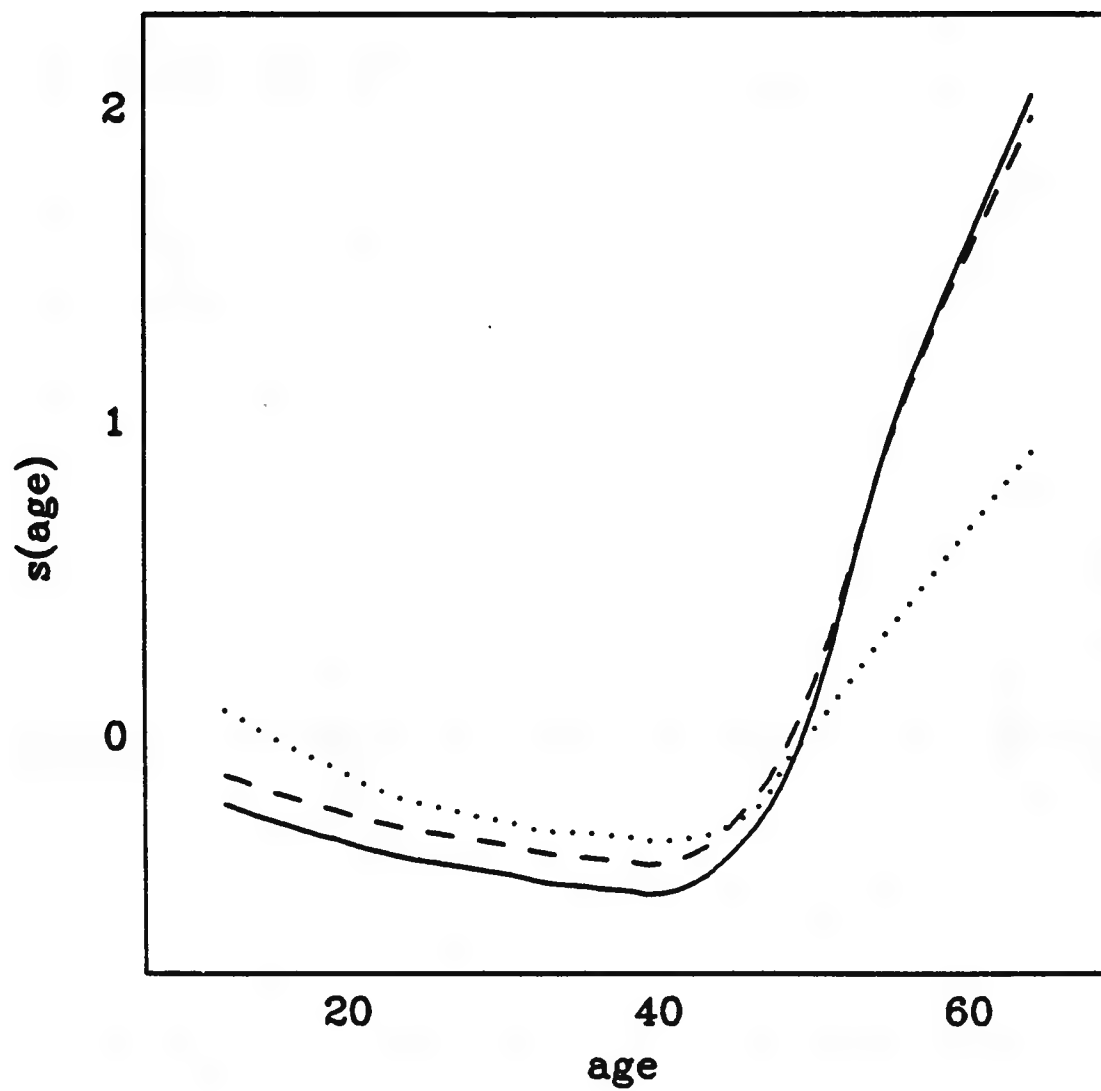
Solid line: Weighted smooth: no outlier  
*Broken Line: Unweighted smooth, no outlier*



**Figure (22)**

Solid line: Weighted smooth, no outlier

Broken and dotted lines: Weighted and unweighted smooths with outlier



## 5. Asymptotic Results.

### 5.1. Introduction

Since local likelihood estimates in the exponential family are maximum likelihood estimates calculated locally, it is not surprising that they enjoy (in some sense) the optimality properties of m.l.e.'s. Tibshirani(1984) modifies McCullagh's (1983) results for generalized linear models to establish such results for l.l.e.'s in the exponential family. We summarize these results below. Similar results should be obtainable for the general (non-exponential) i.i.d case and for the Cox model; we postulate results for the latter at the end of this section.

### 5.2. Consistency and Efficiency of LLE's in the Exponential Family

Consider initially a sample  $\{(x_1, y_1), \dots, (x_n, y_n)\}$  containing an observation at a point  $X = x_0$ . We shall investigate asymptotic properties of the LLE at  $x_0$ . We assume that  $y_1, y_2, \dots, y_n$  are independent realizations of random variables  $Y_1, \dots, Y_n$ , having density

$$Y_i \sim k(y_i) \exp\{y_i \theta_i - b(\theta_i) - c(y_i, \sigma)\} \quad (54)$$

where  $\theta_i = s(x_i)$ . Let  $k_n$  be the number of points in the neighborhood  $N_0^n$  used for estimating  $s(x_0)$ . Assume that as  $n \rightarrow \infty$ ,  $k_n \rightarrow \infty$ , but the neighborhood shrinks so that  $\max_{\{i,j \in N_0^n\}} |x_i - x_j| = o(k_n^{-1/2})$ . We argue below that for estimation of the slope and intercept of the line tangent to  $s(\cdot)$  at  $x_0$ , the LLE is consistent and asymptotically normal, and has the efficiency of a MLE based on sample size  $k_n$ . This implies that, for estimation of  $s(x_0)$ , the LLE has minimum asymptotic mean squared error among all estimates based on  $k_n$  observations.

Letting  $X = (1, x)$ , the score function for the local likelihood at  $x_0$  is

$$U_\beta = X'W(Y - u(\beta)) \quad (55)$$

where  $u(\beta) = b'(X\beta)$ , and  $W = \text{Diag}\{I(i \in N_0^n)\}_{n \times n}$ .

Let  $u(\beta) = E(Y) = b'(\theta)$ ,  $i_\beta = \text{Var}(U_\beta)$  and  $\beta = (\beta_1, \beta_2)$  be the coefficients of the line tangent to  $s(\cdot)$  at  $x_0$ , i.e.  $\beta_2 = s'(x_0)$  and  $\beta_1 = s(x_0) - \beta_2 x_0$ . Then under regularity conditions on  $i_\beta$  and the third moment of  $Y$ , and smoothness constraints on  $s(\cdot)$  and  $b(\cdot)$ , we have

$$k_n^{-1/2} U_\beta \sim \mathcal{N}_2(0, \sigma^2 i_\beta / k_n) + O_p(k_n^{-1/2}) \quad (56)$$

$$E(\hat{\beta} - \beta) = O(k_n^{-1}) \quad (57)$$

and

$$k_n^{1/2}(\hat{\beta} - \beta) \sim \mathcal{N}_2(0, k_n \sigma^2 \mathfrak{s}_{\beta}^{-1}) + O_p(k_n^{-1/2}) \quad (58)$$

These imply the following results for the local likelihood estimate  $\hat{s}(x_0) = \hat{\beta}_1 + \hat{\beta}_2 x_0$ :

$$E(\hat{s}(x_0) - s(x_0)) = O(k_n^{-1}) \quad (59)$$

and

$$k_n^{1/2}(\hat{s}(x_0) - s(x_0)) \sim \mathcal{N}(0, k_n \sigma^2 A) + O_p(k_n^{-1/2}) \quad (60)$$

where  $A = (1 \ x_0) \mathfrak{s}_{\beta}^{-1} (1 \ x_0)^t$ .

### 5.3. Some Remarks

- As mentioned previously, we proved the above results by extending McCullagh's (1983) results for generalized linear models, McCullagh starts with the score equation  $D^t V^{-1}(Y - u(\beta)) = 0$  where  $D = du/d\beta$  and  $V = \text{Cov}(Y)\sigma^2$ . (This reduces to the form  $X^t(Y - u(\beta))$  when the link function is such that  $\theta = X\beta$ ). From this he proves consistency and asymptotic normality of the estimate  $\hat{\beta}$ . Also, he notes that to obtain the asymptotic results, it is not necessary to assume a form for the likelihood: one need only assume that the score equation has the form  $D^t V^{-1}(Y - u(\beta))$ . Since this equation only depends on the first two moments of  $Y$ , there can be more than one likelihood giving the same score equation. McCullagh calls any likelihood giving this score function a "quasi-likelihood". If  $Y$  is in the exponential family and the log-likelihood is linear in  $\eta$ , then the likelihood and quasi-likelihood correspond. In other cases, there can be more than one likelihood resulting in the same quasi-likelihood. In this event, the quasi-likelihood estimate may not equal the MLE, but it is still consistently and efficiently estimates the true parameter. According to McCullagh, "quasi-likelihood" estimation could be useful in a situation in which one isn't willing to assume a specific form for the likelihood, but is willing to specify a relationship between the mean and variance. The same phenomenon is true in the local likelihood model— we need only assume that the local score equation has form (55), and the results still hold.
- The results above assumed that the maximum distance between any two points in a neighborhood goes down at the rate  $o(k_n^{-1/2})$ . In the local likelihood procedure, the span is chosen to minimize an Akaike-type criterion. In principle, then, one should show that selecting the span in this way results in the correct order of shrinkage of the neighborhood. We haven't pursued this, however,
- We have established convergence results for the estimate of a single value of the smooth function. With considerably more work, one could presumably show convergence of the entire estimated function to a Gaussian process.

#### 5.4. Asymptotics for the Proportional Hazards Model

In this section, we conjecture an asymptotic result for the local likelihood procedure in the proportional hazards model.

Suppose  $n$  items are placed on test and give rise to (possibly censored) observation times  $\{y_1, y_2, \dots, y_n\}$  with associated (fixed) covariates  $\{x_1, x_2, \dots, x_n\}$ . Let  $\delta_i = 0$  if  $y_i$  is censored and 1 if  $y_i$  is uncensored, and following Tsiatis(1980), we assume that the triples  $(y_i, x_i, \delta_i)$  are i.i.d. Let  $D$  be the set of indices of the failures among the  $y_i$ 's. To facilitate construction of a partial likelihood, we will make the usual assumption that the censoring mechanism is non-informative (see Kalbfleisch and Prentice(1980)).

Under the model

$$\lambda(t | x) = \lambda_0(t) \exp(x\beta) \quad (61)$$

the partial likelihood is

$$PL = \prod_{l \in D} \frac{e^{\beta x_l}}{\sum_{j \in R_l} e^{\beta x_j}} \quad (62)$$

and the score function is

$$u(\beta) = \sum_{l \in D} \left( x_l - \frac{\sum_{j \in R_l} x_j e^{\beta x_j}}{\sum_{j \in R_l} e^{\beta x_j}} \right) \quad (63)$$

Tsiatis shows that there exist a consistent root  $\hat{\beta}$  of the score equation such that

$$n^{1/2}(\hat{\beta} - \beta) \rightarrow \mathcal{N}(0, v) \quad (64)$$

where  $v = \int_0^{T_0} -dQ \text{Var}(z | R(t))$ ,  $Q(t) = P(t \geq t, \delta = 1)$ , and  $T_0$  is an upper bound on  $Y$ .

In the local likelihood framework, we assume that the hazard has the form

$$\lambda(t | x) = \lambda_0(t) \exp(s(x)) \quad (65)$$

where  $s(x)$  is some smooth function of  $x$ . The derivative  $s'(x_0)$  at some fixed point  $x_0$  is estimated by  $\hat{\beta}_0$  maximizing the local partial likelihood

$$PL_0 = \prod_{l \in D \cap N_0^n} \frac{e^{\beta_0 x_l}}{\sum_{j \in R_l \cap N_0^n} e^{\beta_0 x_j}} \quad (66)$$

The local score equation is

$$u_0(\beta_0) = \sum_{l \in D \cap N_0^n} \left( x_l - \frac{\sum_{j \in R_l \cap N_0^n} x_j e^{\beta_0 x_j}}{\sum_{j \in R_l \cap N_0^n} e^{\beta_0 x_j}} \right) \quad (67)$$

As in the exponential family case, we assume that as  $n \rightarrow \infty$ ,  $k_n \rightarrow \infty$  and  $\max_{\{i, j \in N_0^n\}} |x_i - x_j| = o(k_n^{-1/2})$ . A reasonable conjecture, under regularity conditions on  $s(\cdot)$ , is that the local score equation has a consistent root  $\hat{\beta}_0$ , and asymptotically

$$k_n^{1/2}(\hat{\beta}_0 - s'(x_0)) \rightarrow \mathcal{N}(0, v) \quad (68)$$

where  $v$  is defined above. Fixing  $\hat{s}(x') = s(x') = 0$  for some  $x'$ , a result like (68) could also be obtained for  $\hat{s}(x_0)$ . This would require a convergence proof for the integral estimator  $s(x_0) = \int_{x'}^{x_0} s'(t)dt$ , and hence consideration of the simultaneous estimation of  $s(\cdot)$  at  $x_1, x_2, \dots, x_n$ . We will not attempt to prove these results; the simpler case treated by Tsiatis is quite involved. Recently, more general results for the proportional hazards model (not requiring that the triples  $(y_i, x_i, \delta_i)$  be i.i.d) have been obtained using a martingale approach by Anderson and Gill (1982). A modification of those results to local likelihood estimation should also be possible.

As for efficiency considerations, Efron(1977) and Oakes(1977) show that the partial likelihood estimate has good asymptotic efficiency relative to the full likelihood estimate if the family of hazard models is rich enough. Similar results may be obtainable for the local partial likelihood estimate.

## 6. Degrees of Freedom and AIC Approximations.

In the data analyses of the preceding sections, we have used the formula *degrees of freedom*  $\approx \text{trace}(P)$ . Here we state more precisely the result that is shown in Tibshirani(1984) and discuss a small simulation study to check its accuracy. Finally, we justify the use of the *AIC* is choosing the span. As before, we concentrate on the exponential family case, although our simulations suggest that similar results might be obtainable in the proportional hazards model as well.

We assume that the  $Y_i$ 's are independent with density of the exponential family form

$$g_{\theta_i}(y_i) = \exp(y_i\theta_i - b(\theta_i) - c(y_i, \sigma)) \quad (69)$$

with respect to some carrier measure. The scale parameter  $\sigma$  plays no special role and is assumed to be 1 for simplicity. Let  $k_{\theta}(\mathbf{y}) = \prod_{i=1}^n g_{\theta_i}(y_i)$ , and let  $\mathbf{b}(\theta) = (b(\theta_1), b(\theta_2), \dots, b(\theta_n))$ . It will be convenient to instead index  $k(\mathbf{y})$  by the expectation parameter  $\mu = E_{\theta}\mathbf{y} = \mathbf{b}'(\theta)$ . The deviance between  $\mathbf{y}$  and  $\hat{\mathbf{y}}$  is defined by

$$D(\mathbf{y}, \hat{\mathbf{y}}) = 2[\log k_{\mathbf{y}}(\mathbf{y}) - \log k_{\hat{\mathbf{y}}}(\mathbf{y})] \quad (70)$$

A single covariate  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  is available.

Central in our discussion will be the "smoother matrix"  $P$  "corresponding" to a local likelihood fit  $\hat{\mathbf{y}}$ . Given a linear covariate vector  $\mathbf{x}$ , the running lines smooth of a data vector  $\mathbf{y}$ , based on  $\mathbf{x}$ , can be written as  $\hat{\mathbf{y}} = P\mathbf{y}$ . We call  $P$  a "smoother matrix"; it will depend on  $\mathbf{x}$  and on the span of the smoother. Given a running lines smoothing algorithm, it is easy to produce  $P$ —the  $i$ th row of  $P$  is the output of the smoother applied to the  $i$ th unit vector.

For Gaussian data, the local likelihood fit is simply  $\hat{\mathbf{y}} = P\mathbf{y}$ . For non-Gaussian likelihoods, the matrix  $P$  doesn't enter in the local likelihood estimation procedure explicitly, but its *trace* turns out to be important nonetheless.

Now let  $\hat{\mathbf{y}}_1$  and  $\hat{\mathbf{y}}_2$  be two local likelihood fits, based on  $\mathbf{x}$ , with corresponding smoother matrices  $P_1$  and  $P_2$ . Suppose that the two smooths produce the same fit on the average, i.e.  $E(\theta_{\hat{\mathbf{y}}_1}) = E(\theta_{\hat{\mathbf{y}}_2})$ , where in obvious notation,  $\theta_{\mathbf{y}} = \mathbf{b}'^{-1}(\mathbf{y})$ . Then it is shown in Tibshirani (1984) that

$$E(D(\mathbf{y}, \hat{\mathbf{y}}_1) - D(\mathbf{y}, \hat{\mathbf{y}}_2)) \approx \text{trace}(P_2) - \text{trace}(P_1) \quad (71)$$

This result is exact for the Gaussian case, where the deviance is the residual sum of squares, and approximate for other exponential family likelihoods. We describe in the next section a simulation study to assess the accuracy of this approximation,

Note how (71) generalizes the standard result for global linear fitting. Consider a normal linear regression model, with variance equal to 1. Let  $\hat{\mathbf{y}}_1$  and  $\hat{\mathbf{y}}_2$  represent the fitted vectors corresponding to nested subspaces of rank  $p_1$  and  $p_2$ , with  $p_1 < p_2$ . Then if the true mean response lies in the smaller space, then the decrease in residual sum of squares due to fitting the

larger model is  $\chi^2_{p_2-p_1}$ . Result (71) expresses this in expectation, for Kullback-Leibler distance is Euclidean distance in the Gaussian model, and  $P_1$  and  $P_2$  are the matrices that project onto the corresponding subspaces (recall that for a projection matrix,  $\text{trace}(P) = \text{rank}(P)$ ). For linear fitting in other exponential family models (GLM's), result (71) expresses Wald's theorem in expectation.

### 6.1. Degrees of Freedom Simulations

Table 5 shows the results of a modest simulation study designed to check the accuracy of the formula  $E(D(\mathbf{y}, \hat{\mathbf{y}}_1) - D(\mathbf{y}, \hat{\mathbf{y}}_2)) = \text{trace}(P_2) - \text{trace}(P_1)$ .

**Table 5. Results of Degrees of Freedom Simulation**

Entries in Lines (2)–(5) are mean(variance) of deviance decrease

Source	Span				
	.3	.4	.5	.6	.7
(1) $\text{Trace}(P) - 1$	4.09	3.32	2.65	2.34	2.16
(2) Scatterplot Smooth( $\mathbf{y}$ normal)	4.14(10.00)	3.39(7.75)	2.61(6.03)	2.31(5.08)	2.09(4.32)
(3) Scatterplot Smooth( $\mathbf{y}$ uniform)	4.19(10.06)	3.46(8.50)	2.77(6.52)	2.41(5.79)	2.21(4.99)
(4) Logistic Model(constant vs smooth)	4.34(13.47)	3.40(11.62)	2.72(9.12)	2.23(7.51)	2.17(6.28)
(5) Logistic Model(linear vs smooth)	3.29(11.71)	2.25(8.25)	1.63(6.21)	1.29(4.58)	1.12(2.89)
(6) Cox Model(no censoring)	5.58(13.37)	4.24(8.99)	3.63(7.52)	3.12(6.25)	2.71(5.48)
(7) Cox Model(40% censoring)	5.36(13.54)	4.16(9.04)	3.62(6.98)	3.13(5.86)	2.73(5.20)

The numbers in the table were obtained as follows. 100  $\mathbf{z}$  values were generated from  $\mathcal{N}(0, 1)$  and fixed for the entire table. Given these  $\mathbf{z}$  values, we constructed the running lines smoother matrices for the indicated spans, and the trace of each matrix (minus 1) is shown in line (1).

Consider for example the entry 4.09 in the top left hand corner. According to the preceding derivation, this should be the expected decrease in deviance due to fitting a local likelihood model with that span .3 versus a model with only a constant.

To obtain line (2), we generated 100  $\mathbf{y}_i$ 's from  $\mathcal{N}(0, 1)$  and computed  $R(\mathbf{y}, \bar{\mathbf{y}}_1) - R(\mathbf{y}, \hat{\mathbf{y}})$ ,  $\hat{\mathbf{y}}$  being the fit from a scatterplot smoother ( $\hat{\mathbf{y}} = P\mathbf{y}$ ) with span as shown. Line(2) shows the mean and variance from 500 such repetitions of this process.

Line (3) was obtained in same way as line (2), except that the  $y_i$ 's were generated from uniform  $(-\sqrt{3}, \sqrt{3})$ , the range chosen so that  $Var(y_i) = 1$ .

To obtain line(4), we generated 100  $y_i$ 's from *binomial*(1, 1/2) and fit a smooth logistic model with spans of .3 to .7. The numbers show the mean and variance of  $D(y, \hat{y}_1) - D(y, \hat{y})$  over 500 repetitions.

Line (5) was generated in a similar fashion as line (4), showing instead the mean and variance of  $D(y, \hat{y}_1) - D(y, \hat{y})$ ,  $\hat{y}_1$  being the linear logistic fit, with  $y_i$  generated from a linear logistic model,  $P(y_i = 1 | x) = e^{2x} / (1 + e^{2x})$ .

Lines (6) and (7) show simulation results for the Cox model. 100  $y$  values were generated according to  $y = \exp(1 + \epsilon)$ , where  $\epsilon$  had an extreme value distribution. This corresponds to a constant hazard (exponential) model. For line (6), no censoring was applied. For line (7), censoring variables  $c_i$  were generated from  $e^{6u}$ ,  $u \sim U(0, 1)$ . This produced a censoring rate of about 40%. A smooth Cox model was fit and the quantity  $-2 \log L(\text{null model}) - (-2 \log L(\text{smooth}))$  was computed. Lines (6) and (7) show the mean and variance of this quantity over 500 repetitions.

Note that for all the models, a span of 100% gives either exactly or asymptotically a mean value of 1 (by Wald's theorem), and  $trace(P) - 1$  is also equal to 1.

The results give fairly strong support to the approximation  $E(D(y, \hat{y}_1) - D(y, \hat{y}_2)) = trace(P_2) - trace(P_1)$ . Lines (2) and (3) agree well with (1), not surprising since the approximation is exact for scatterplot smoothers. Line (4) also is in good agreement, with a small upward bias for smaller spans. Line (5) should be 1 less than line (1), (since the global linear fit uses 2 degrees of freedom) and the results indicate that. For the Cox model, the *trace* formula is biased downward, especially in the lower spans.

The variance results are a little unsettling. The variance to mean ratio is often greater than 2 (the ratio for a chi-square variate), especially for the non-gaussian models.

We conclude from these simulations that the approximation  $E(D(y, \hat{y}_1) - D(y, \hat{y}_2)) = trace(P_2) - trace(P_1)$  is satisfactory as a rough rule of thumb, for the gaussian and logistic models. We do note, however, that the distribution of this decrease is more spread out than a chi-square variate with the corresponding degrees of freedom, so that tests based on the percentile points will be too liberal. The downward bias of *trace*( $P$ ) the Cox model does not cause a serious problem in the *AIC* formula  $-2 \log L + 2trace(P)$ , since the bias will tend to cancel out in comparing the *AIC* for two spans. In assessing the significance of a Cox smooth in real examples, however, we find the mean decrease by simulation as in Table 5. Indeed, for any of the models, it may be preferable to get "exactly" the distribution of the decrease by simulation, for these simulations are not expensive. Finally, we would like to acknowledge that the idea of finding degrees of freedom by simulation for a scatterplot smoother was first suggested to the author by Arthur Owen.

### 6.2. Akaike's Information Criterion(AIC) For Span Selection

Using the results of the previous section, we show in this section that it's reasonable to use an *AIC* criterion to choose the span in the local likelihood estimation procedure.

Let's briefly review the *AIC* for a parametric model. Given a model  $k_{\mu}$ , suppose we can choose among maximum likelihood estimates  $\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_l$  based on  $p_1, p_2, \dots, p_l$  degrees of freedom respectively. Suppose also that each model can be considered a sub-model of a true model  $k_{\mu_0}$ . Then Akaike's information criterion (*AIC*) (Akaike 1973) specifies that we should choose the estimated model that minimizes

$$AIC = -2 \log k_{\hat{\mu}_i}(\hat{y}) + 2p_i \quad (72)$$

Akaike derived the *AIC* by showing that  $E(AIC) \approx E(D(\mu_0, \hat{\mu}_i)) + \text{constant}$ . Hence the model that minimizes *AIC* approximately minimizes the expected Kullback-Leibler distance from the true model.

From the form of the *AIC*, it is clear that it attempts to trade-off goodness of fit of the estimated model with its complexity. Not surprisingly, it turns out to be identical to Mallows's  $C_p$  in the linear regression setting and asymptotically equivalent to the cross-validated likelihood technique in general (see Stone (1977) for these results).

In the local likelihood procedure, we propose choice of the span parameter  $s$  to minimize

$$AIC = -2 \log k_{\hat{y}(s)}(\hat{y}) + 2 \text{trace}(P(s)) \quad (73)$$

where  $P(s)$  denotes the smoother matrix producing running lines fits with span  $s$ , and  $\hat{y}(s)$  denotes the corresponding fitted values. This makes sense intuitively: as the span  $s$  increases, the first term will (generally) increase but the degrees of freedom  $\text{trace}(P(s))$  will decrease. Hence the *AIC* will trade off lack of fit with complexity of the smooth.

In Tibshirani (1984), we show that the *AIC* approximately equals a measure of expected distance to the true model. The logic of the derivation follows that of Akaike (1973). Consider the exponential family set-up of the previous section. Using the notation of that section, we let  $P$  be a running lines smoother matrix corresponding to some span. Then we show that

$$E(D(\mu_0, \hat{y})) \approx -n + 2 \text{trace}(P) \quad (74)$$

Noting that  $D(\hat{y}, \hat{y}) = -2 \log k_{\hat{y}}(\hat{y}) + \text{constant}$  and also that  $n$  is constant for all spans, we see that the fit  $\hat{y}$  minimizing the *AIC* criterion (73), minimizes an estimate of distance to the true model  $k_{\mu_0}$ .

### Software

Fortran programs that compute local likelihood estimates in the Cox model and in generalized linear models, are available from the author.

### ACKNOWLEDGMENTS

The author wishes to thank Trevor Hastie for his contributions to this work, including first suggesting the use of local likelihood in generalized linear models, work on the logistic model, and coining of the name "local likelihood". Thanks also to Bradley Efron, Jerome Friedman, and Paul Switzer for their helpful comments, and Art Owen for his ideas on finding degrees of freedom by simulation.

This work formed part of the author's Phd dissertation at the Department of Statistics, Stanford University, supervised by Professor Bradley Efron. The author was supported by an Natural Sciences and Engineering Research Council of Canada Post-Graduate Fellowship, by the Department of Energy under contracts DE-AC03-76SF00515 and DE-AT03-81-ER10843, by the Office of Naval Research under contract ONR N00014-81-K-0340, and by the U.S. Army Research Office under contract DAAG22-82-K-0056.

## REFERENCES

- Akaike, H. (1973) Information theory and an extension of the entropy maximization principle. 2nd International Symposium on information theory, pp 267-281.
- Anderson, P. and Gill, R. (1982) Cox's regression model for counting processes: a large sample study. *Ann. Stat.* 10, 4, 1100-1120.
- Breiman, L. and Friedman J.H. (1982). Estimating optimal correlations for multiple regression and correlation. Stanford U. tech. rep. Orion 010.
- Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics* 30, 89-99.
- Cain, K. and Lange, N. (1984) Estimating case influence for proportional hazards regression models with censoring. Tech rep 320Z, Dept. of Biostatistics, Dana-Farber Cancer Institute.
- Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, 74 828-836.
- Cox, D.R. (1972). Regression models and life tables. *J. Roy. Stat. Soc. B*, 34, 187-202.
- Cox, D.R. (1975). Partial likelihood. *Biometrika* 62, 269-276.
- Crowley, J. and Hu, M. (1977). Covariance Analysis of heart transplant survival data. *J. Amer. Statist. Assoc.* 72, 27-36.
- Efron, B. (1975). The Geometry of Exponential Families. *Ann. Stat.* 6, No 2, 362-376
- Efron, B. (1977). The Efficiency of Cox's likelihood function for censored data. *J. Amer. Statist. Assoc.* 72, 557-565.
- Efron, B. (1979). Bootstrap Methods: another look at the Jackknife. *Ann. Stat* 7, pp 1-26.
- Efron, B. (1980) Censored data and the bootstrap. *J. Amer. Stat. Assoc.* 76, 312-19.
- Friedman, J.H., and Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* 76, 817-823.
- Friedman, J.H., and Stuetzle, W. (1982). Smoothing of scatterplots. Stanford Univ. technical report - Orion 003.
- Friedman, J.H., and Owen, A. Predictive ACE. In preparation.
- Guttman, I. (1983) *Linear Models*. Wiley, New York.
- Haberman, S. (1976) Generalized Residuals for Log Linear Models. *Proc. of 9th Int. Biost. Conf*, Boston 104-122.
- Hastie, T. (1983). Non-parametric logistic regression. Stanford University Technical report ORION 016.

- Hastie, T. (1984) Discussion of "Graphical Methods for assessing logistic regression models", by Landwehr et al, J. Amer. Stat. Assoc. 79, 61-63.
- Hastie, T., Tibshirani, R., and Owen, A. (1984). Generalized Additive Models. In preparation.
- Kalbfleisch, J.D., and Prentice, R.L. (1980). The Statistical analysis of failure time data. Wiley, New York.
- Kay, R. (1977). Proportional hazard models and the analysis of censored survival data. J. Roy. Stat. Soc. C., 26, 227-237.
- Krasker, W. and Welsch, R. (1982) Efficient bounded influence regression using alternate definitions of sensitivity. J. Amer. Stat. Assoc. 77, 595-605
- Landwehr, J., Pregibon, D. and Shoemaker, A. (1984) Graphical methods for assessing logistic regression models. J. Amer. Stat. Assoc., 79, 61-63.
- McCullagh, P. (1983). Quasi Likelihood Functions. Ann. Stat. 11, 59-67.
- Miller, R.G. and Halpern, J. (1982). Regression with censored data. Biometrika 69, 3, 521-31.
- Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. J. Roy. Stat. Soc. A, 135, 370-384.
- Oakes, D. (1977). The asymptotic information in censored survival data. Biometrika 67, 441-448.
- Peto, R. (1972). Discussion on Professor Cox's paper. J. Roy. Stat. Soc. B, 34, 205-207.
- Prentice, R. and Breslow, N. (1978). Retrospective Studies and Failure Time Models. Biometrika 65, 153-158.
- Simon, G. (1973). Additivity of information in exponential family probability laws. J. Amer. Stat. Assoc. 68, 478-482.
- Stone, M. (1977) As asymptotic choice of model by cross-validation and Akaike's criterion. J. Roy Stat. Soc. B., No 1, vol 7, 44-47.
- Thomas, D. (1983) Non-parametric estimation and tests of fit for dose response relations. Biometrics, Vol 39, No 1, 263-268.
- Tibshirani, R. (1984). Phd Dissertation. Department of Statistics, Stanford University.
- Tsiatis, A. (1981). A large sample study of Cox's regression model. Vol 9, No 1, 93-108.